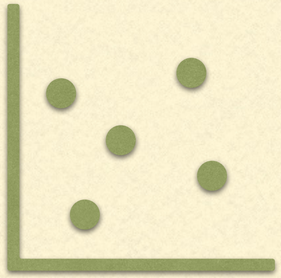
LINEAR REGRESSION, T -TEST, ANOVA,

Nhu L.T. Tran

TERMINOLOGY ALERT - WHAT IS…?

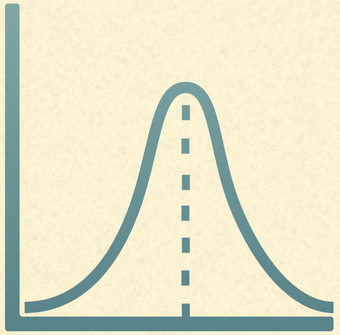
"Household" terms that are often not clearly explained



Regression

Linear model

Variance



Statistical power

SumSq

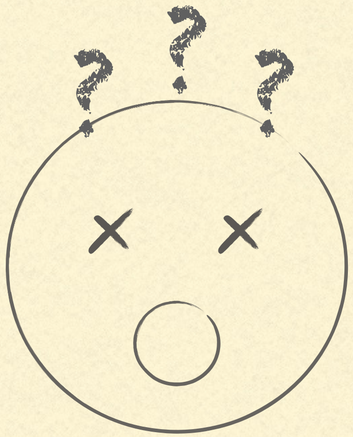
MeanSq

T-test or ANOVA

Standard deviation

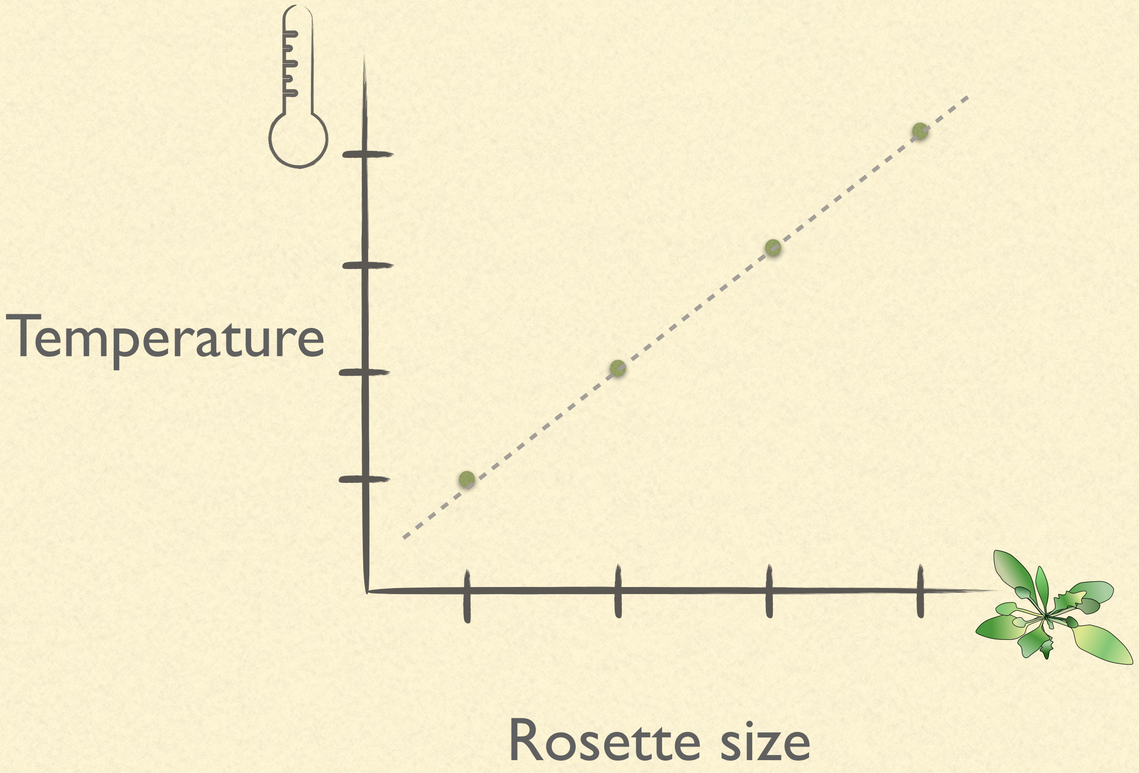
Statistically significant

p-value

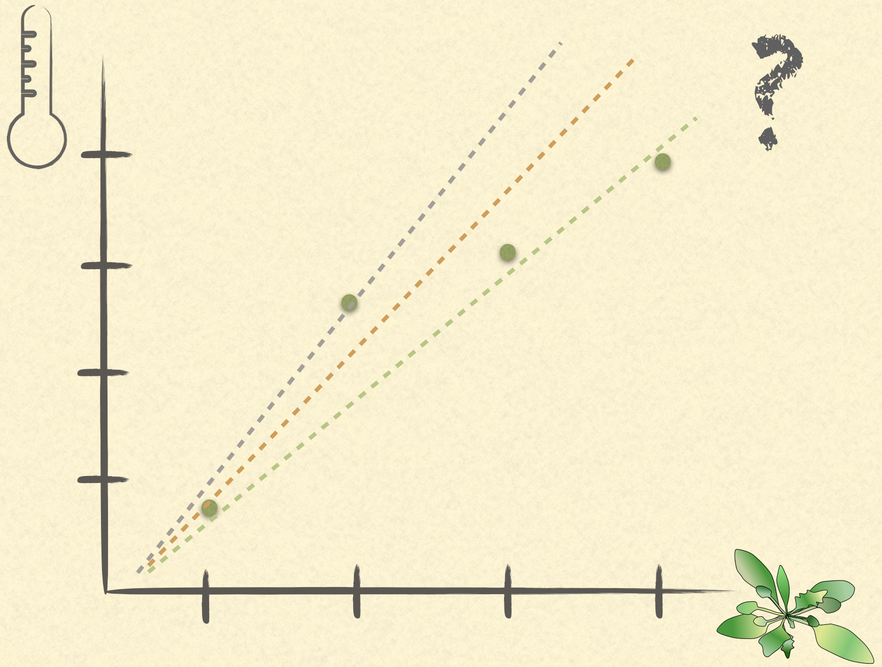


1. LINEAR REGRESSION

Ideal world: Fitting a line across all points

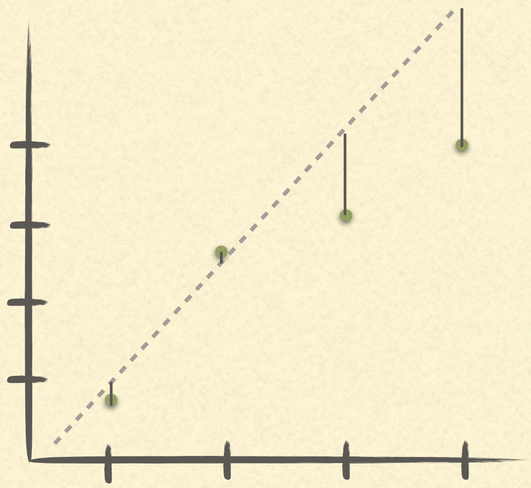


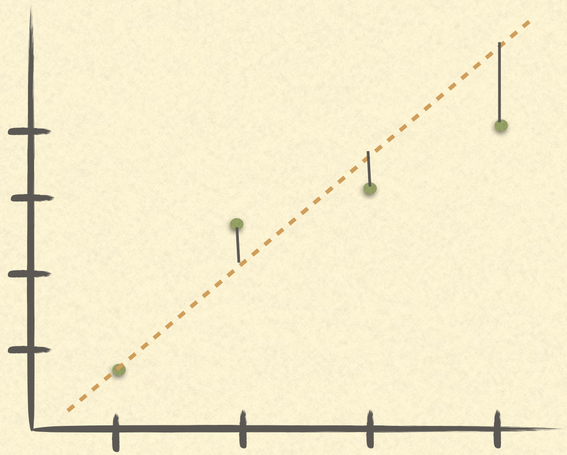
Reality: trying to fit a line across all points…

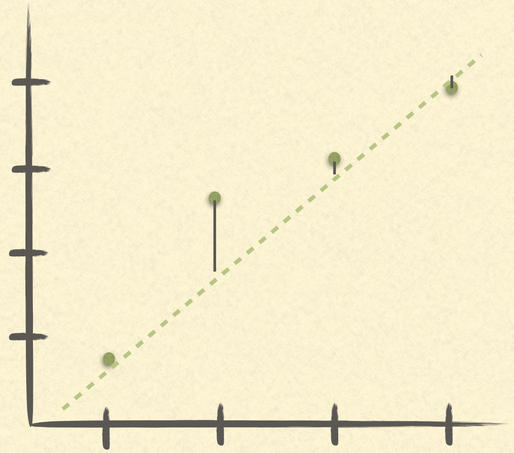


1. LINEAR REGRESSION

Which is the best line?



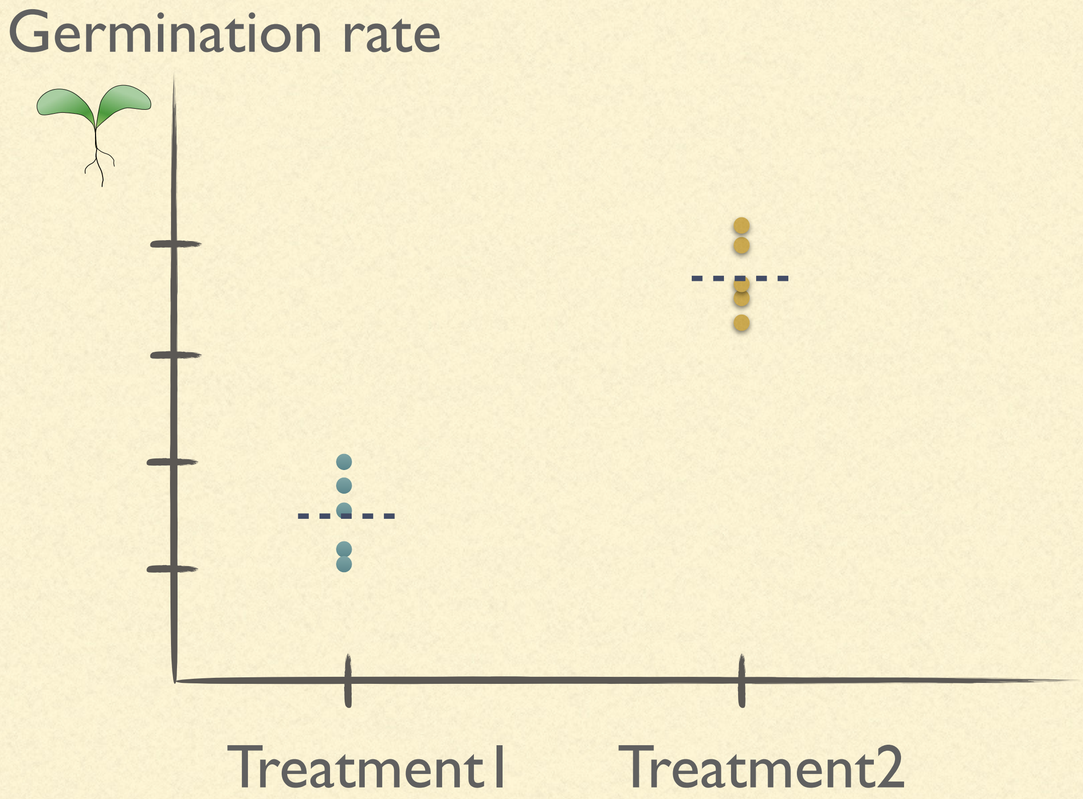




Calculate the (Euclidian) distance from the data points to the line and sum them up (*sum of squares*).
The line with smallest sum of squares is the best fit

This Sum of Squares concept is the basis for understanding many statistical tests…

2. T-TEST



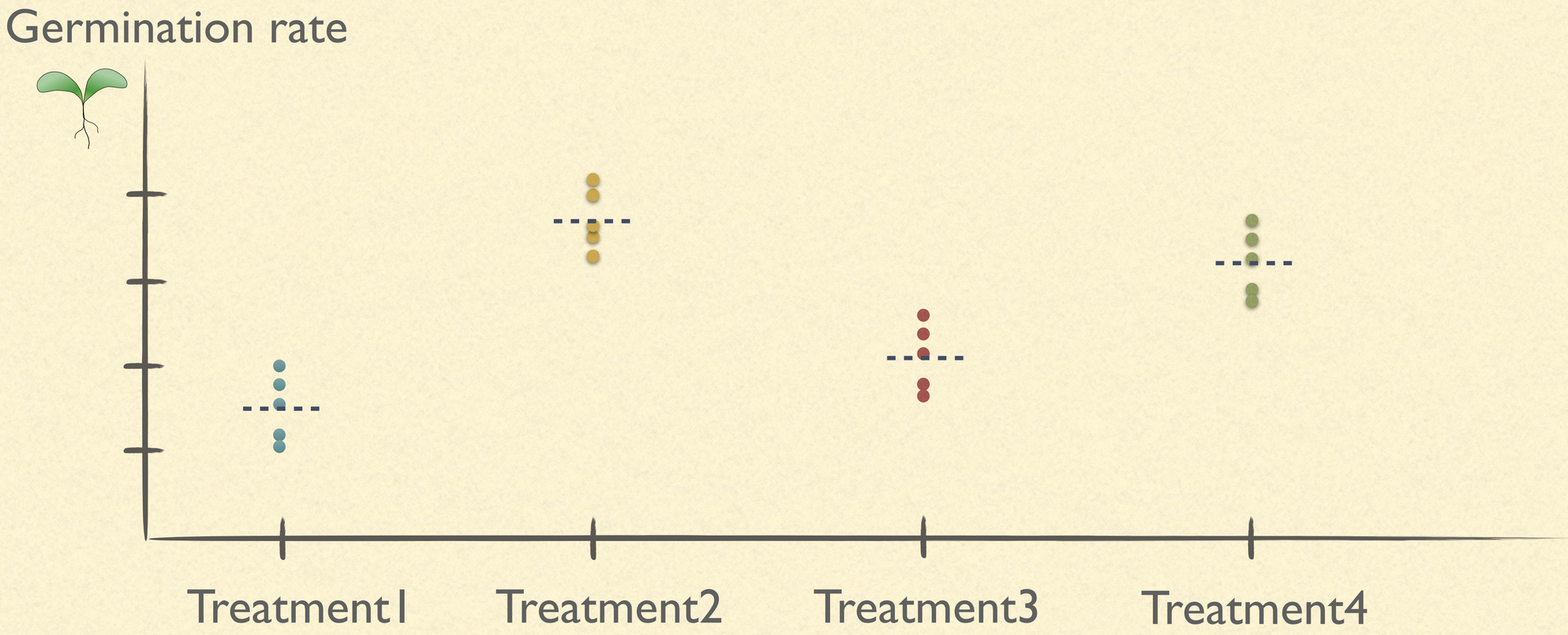
Do the group means differ among 2 treatments?

Seems to be, but you need to prove it statistically

=> using *two-sample T-test*

3. ANOVA = ANALYSIS OF VARIANCE

When you want to do t-test but have more than 2 groups.



3. ANOVA = ANALYSIS OF VARIANCE

Why is it analysis of "variance"?

Because it quantifies the variation in your data. General principle of ANOVA is:

1. Quantify total variation in the data
2. Separate signal (treatment effect) and noise (error)
3. The bigger signal to noise ratio is better

WHEN YOU RUN ANOVA, YOU GET THIS

- What do these Df, SumSq, MeanSq, F value, Pr > F mean??

```
> example_anova1 <- aov(dependent_variable ~ treatment, data = example_dataframe)
> summary(example_anova1) #F-test and ANOVA table
            Df Sum Sq Mean Sq F value  Pr(>F)
treatment    4 11.258  2.8144   6.566 0.00152 **
Residuals   20  8.572  0.4286
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

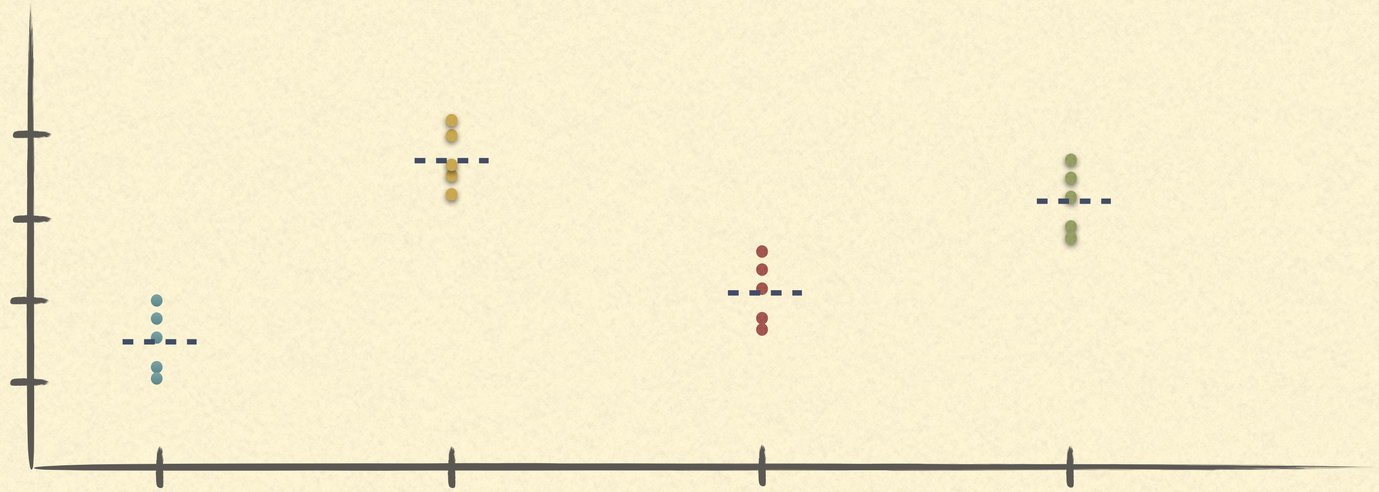
- So we need to get some theory first.

3. THE STEPS BEHIND ANOVA

Step 1: Define null hypothesis

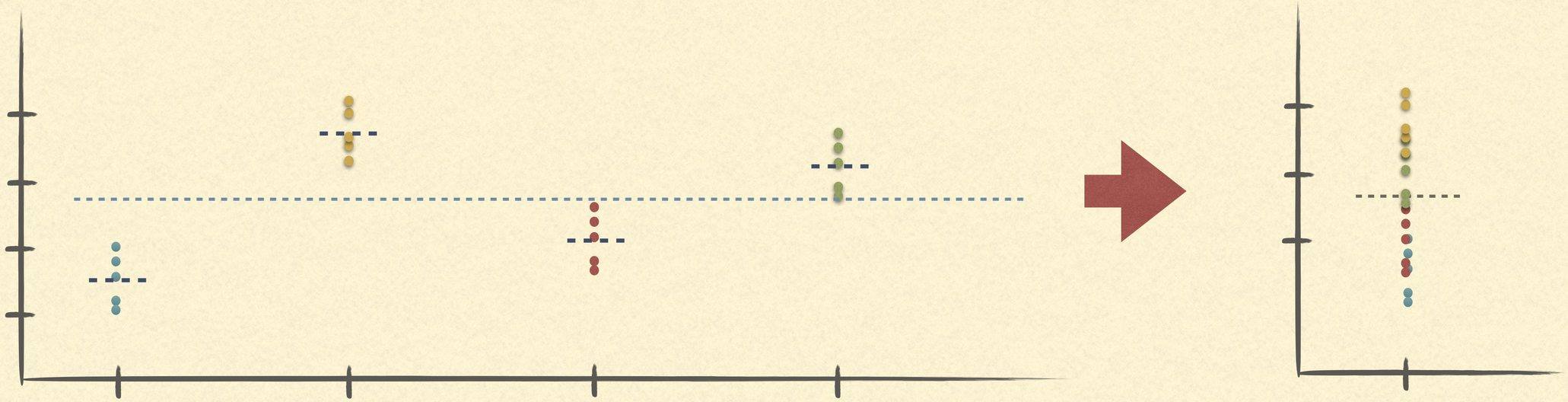
- Null hypothesis (often denoted as **H0**): all group means are equal

- Alternative hypothesis (**H1**): at least one mean is different from the others



3. THE STEPS BEHIND ANOVA

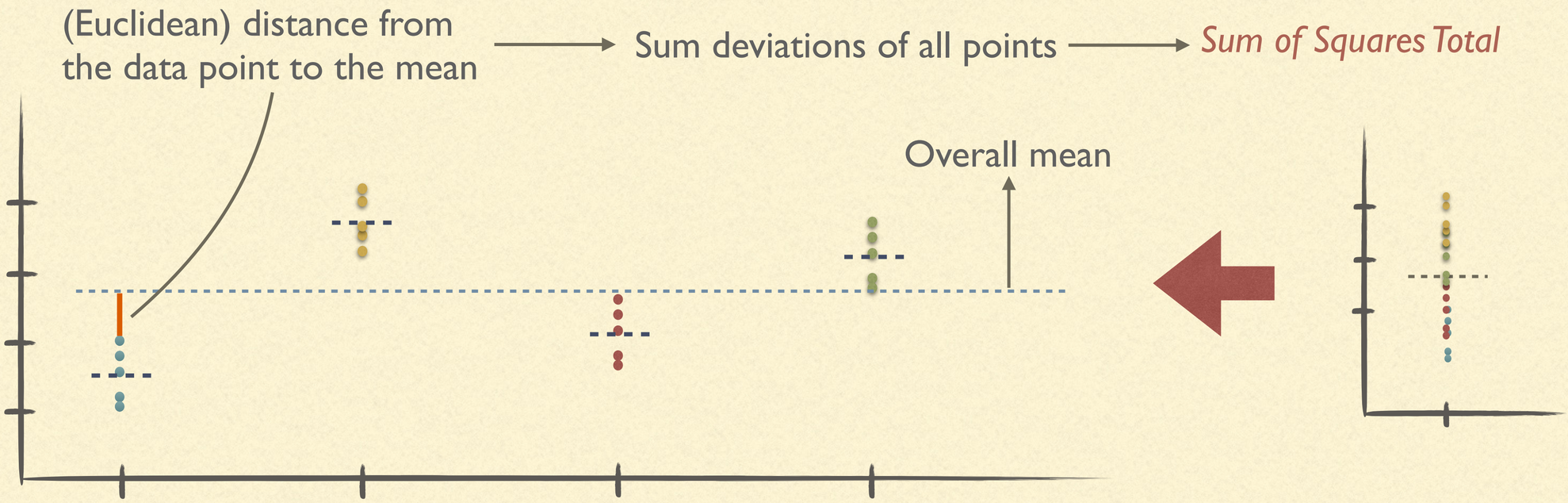
Step 2: Total variation



Overall mean: take all data from all groups together and calculate the mean

3. THE STEPS BEHIND ANOVA

Step 2: Total variation

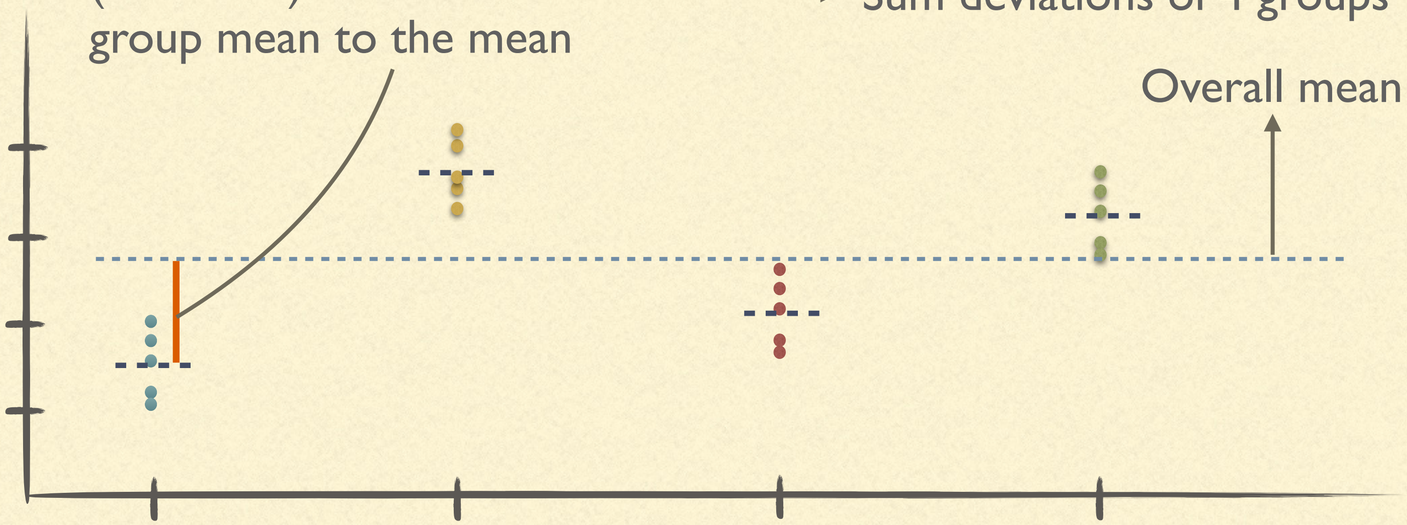


Total variation = deviation of all data points from the overall mean = Variation due to treatments and error

3. THE STEPS BEHIND ANOVA

Step 3: Between-group variation

(Euclidean) distance from the group mean to the mean → Sum deviations of 4 groups → *Sum of Squares of Treatment*

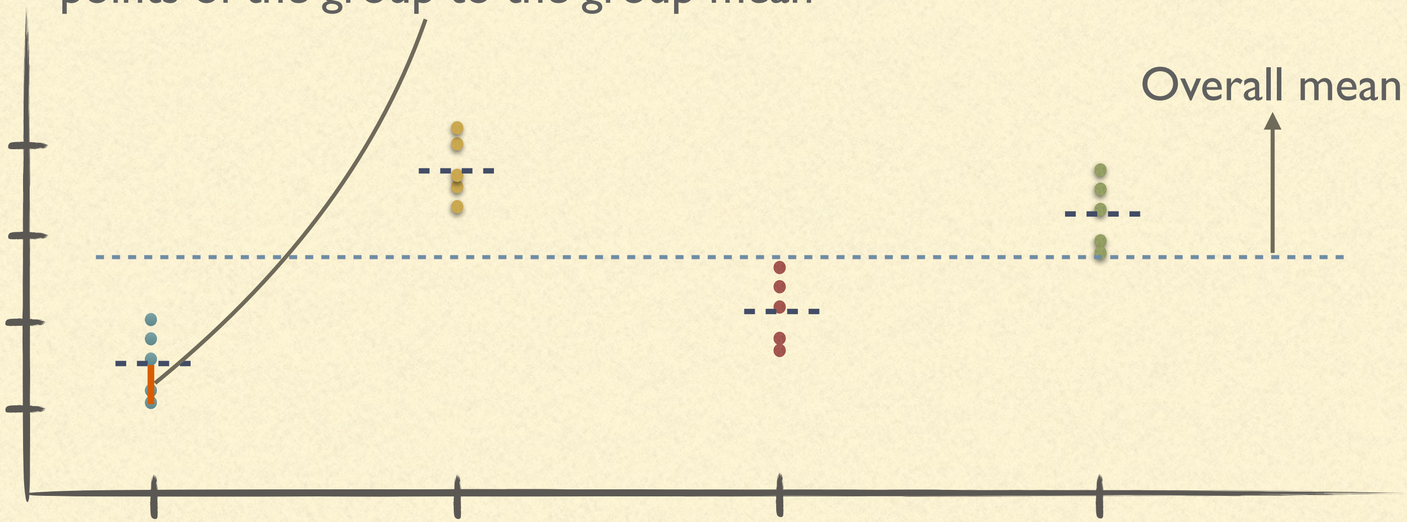


Between-group variation = deviation of group mean from the overall mean = Variation due to treatments

3. THE STEPS BEHIND ANOVA

Step 4: Within-group variation

(Euclidean) distance from the data points of the group to the group mean → Sum all within-group deviations → *Sum of Squares of Error*

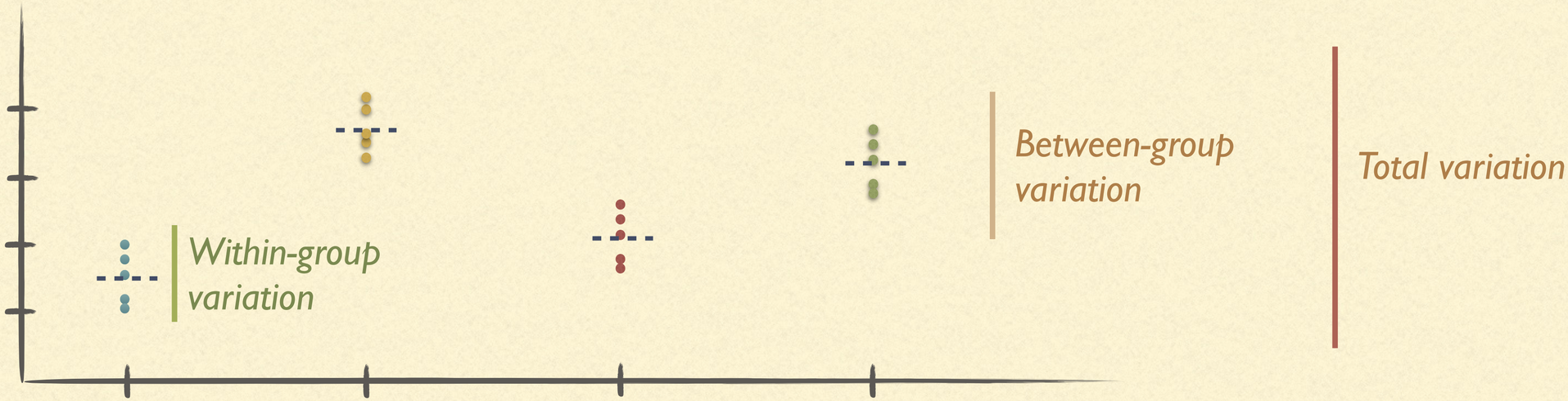


Within-group variation = deviation of data points within the group from that group mean

= Variation due to error (anything that caused variation rather than the treatments)

3. THE STEPS BEHIND ANOVA

Three types of variation



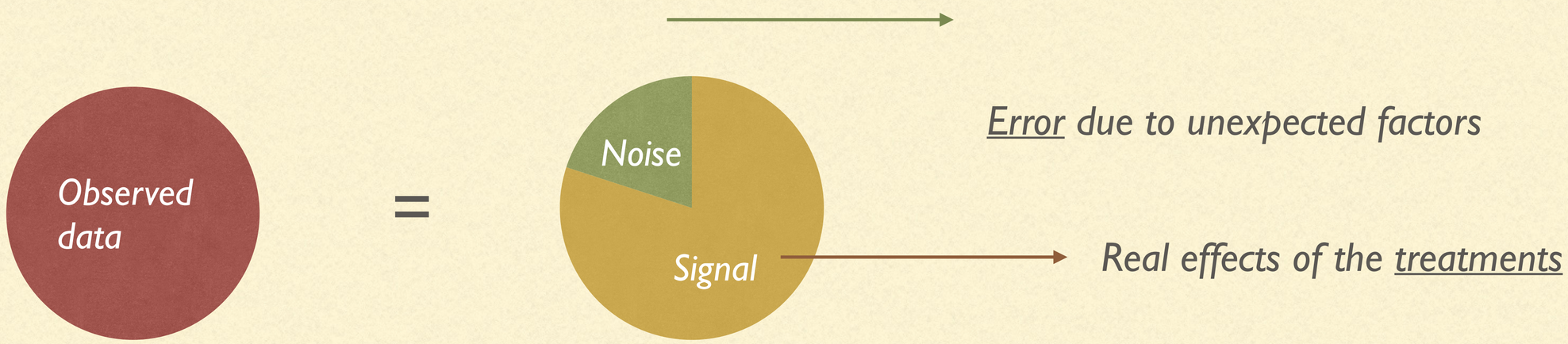
Sum of Squares Total = *Sum of Squares of Treatments* + *Sum of Squares of Error*

3. THE STEPS BEHIND ANOVA

Summary of ANOVA

Sum of Squares Total = *Sum of Squares of Treatments* + *Sum of Squares of Error*

Total variation = *Between-group variation* + *Within-group variation*



4. DEGREE OF FREEDOM (DF)

DF = sample size - 1

Where is it needed?

$$\frac{\textit{Sum of Squares Total}}{\textbf{Df}} = \frac{\textit{Sum of Squares of Treatments}}{\textbf{Df}} + \frac{\textit{Sum of Squares of Error}}{\textbf{Df}}$$

→ *Mean of Squares Total* = *Mean of Squares of Treatments* + *Mean of Squares of Error*

→ *Signal/noise ratio* = *so-called* ***F value*** = *Mean of Squares of Treatments* / *Mean of Squares of Error*

4. DEGREE OF FREEDOM (DF)

But what is it? => DF tells how many **independent** pieces of information went into the calculation of statistics

Guess the traffic light colour. We know it's not green

=> Three possible colours but because of one piece of information (not green),

We two independent pieces of information: Red and Yellow

10
21
?
7

Find the "?". We know the mean is 11

We can easily calculate the "?"

12
21
?
5

Other values changed but we always have the mean = 11

"?" now has to change

=> "?" is dependent on the three other values

Independent values are free to vary, while **dependent** values are dependent on the free values

TEST YOURSELF

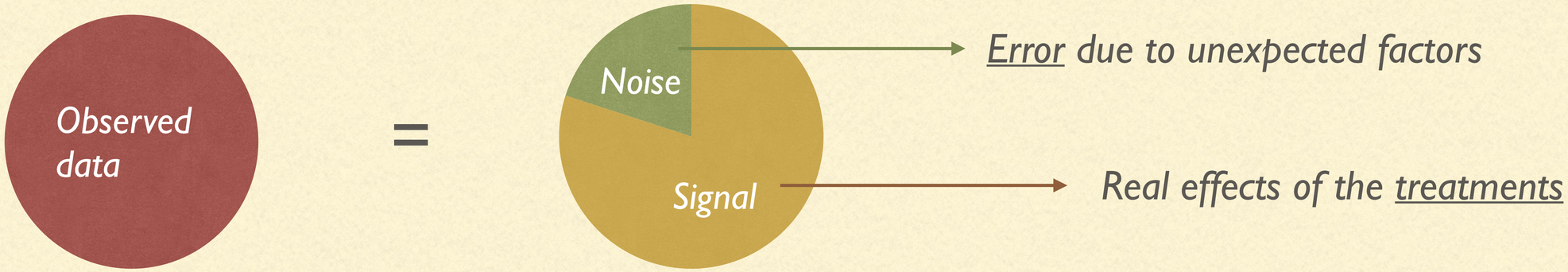
Question

What does it mean that mean squares of between-group effect in ANOVA is 0?

Hint: observed data = overall mean + treatment effect + residual (error)

Sum of Squares Total = *Sum of Squares of Treatments* + *Sum of Squares of Error*

Total variation = *Between-group variation* + *Within-group variation*



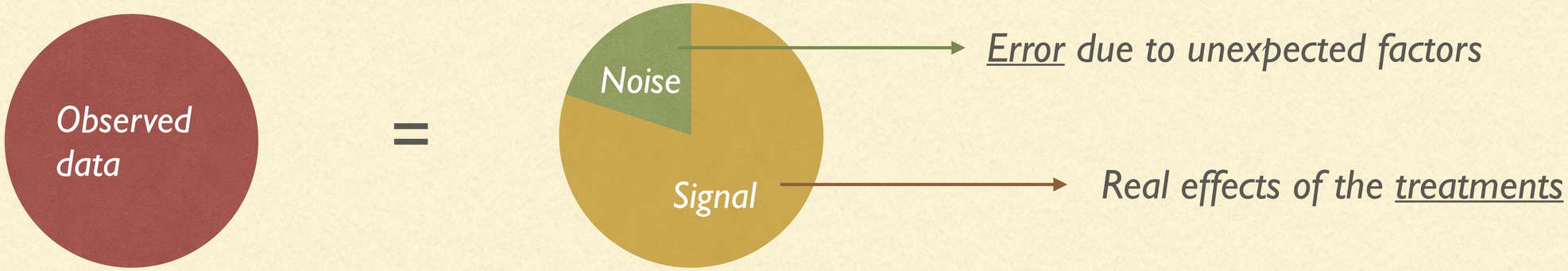
TEST YOURSELF

Answer

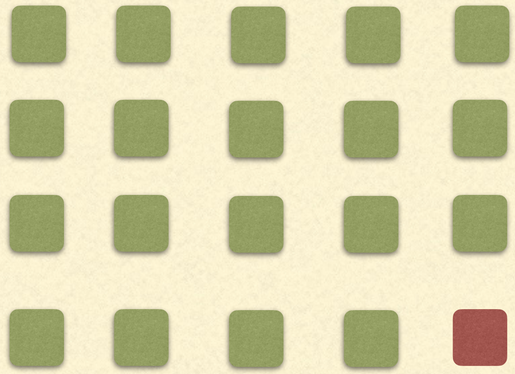
What does it mean that mean squares of between-group effect in ANOVA is 0?

=> *The treatments have no effect in explaining the model, i.e., there is no treatment effect*

Total variation = *Between-group variation* + *Within-group variation*



4. THE P-VALUE



When you perform a statistical test and get good "signal"

=> what's the probability that this is just a random chance (i.e., false positive)?

There can be 5% chance (p value = 0.05) that this result is by chance (no treatment effects). And you're fine with it: you have 95% chance that the effect of your treatment is real.

p value is not any magic number, it's what the community accepts.

You can be happy that 90% chance that the effect is real, but not everyone is happy with it.
Therefore, you can only report when you find something with p value < 0.05

There are studies where people are only happy with p value < 0.01.

4. THE P-VALUE

Remember we set H0 and H1.

- Null hypothesis (often denoted as **H0**): all group means are equal
- Alternative hypothesis (**H1**): at least one mean is different from the others

P < 0.05 => we can "reject the null hypothesis" — meaning the H1 is true.

With this we say there is a significant effect of treatments on the germination rate, accepting that there is 5% chance that this effect occurs due to random chance.

You report it as "the treatment … has statistically significant effect on germination".

TEST YOURSELF

Do you understand this?

```
> example_anova1 <- aov(dependent_variable ~ treatment, data = example_dataframe)
> summary(example_anova1) #F-test and ANOVA table
            Df Sum Sq Mean Sq F value  Pr(>F)
treatment    4 11.258  2.8144   6.566 0.00152 **
Residuals   20  8.572  0.4286
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Which values you want to be large, and which to be small?

TEST YOURSELF

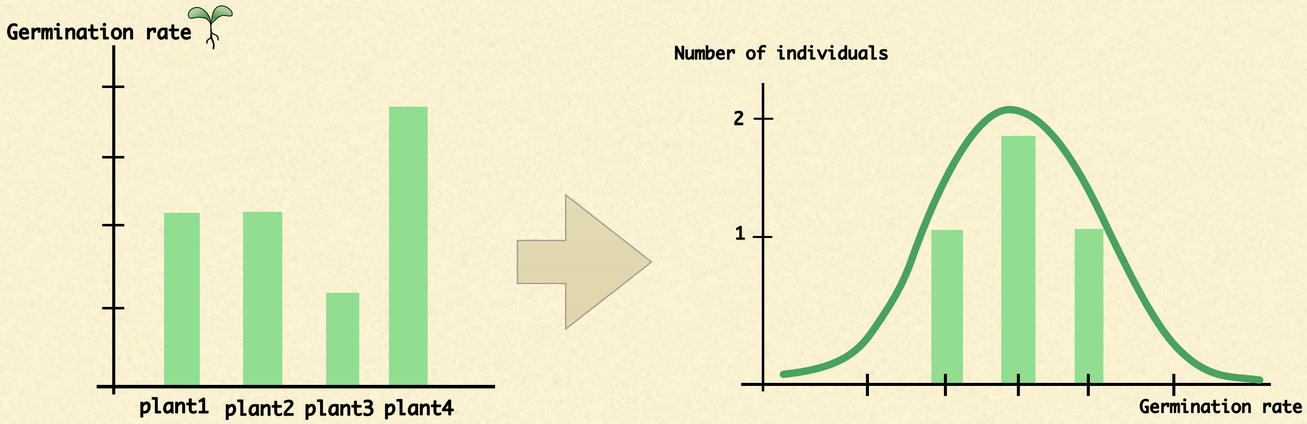
Answer

```
> example_anova1 <- aov(dependent_variable ~ treatment, data = example_dataframe)
> summary(example_anova1) #F-test and ANOVA table
            Df Sum Sq Mean Sq F value  Pr(>F)
treatment    4 11.258  2.8144   6.566 0.00152 **
Residuals   20  8.572  0.4286
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

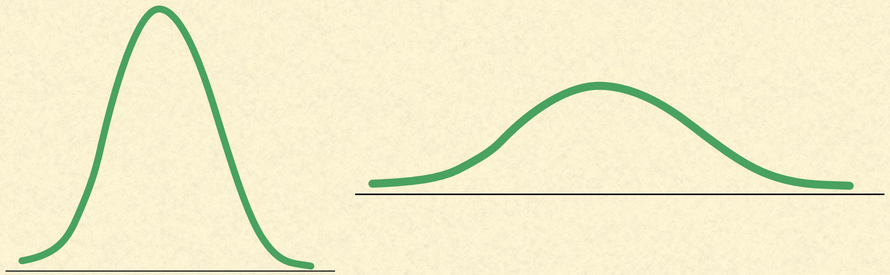
- Values we want to be large: F value (signal to noise ratio); Sum Sq and Mean Sq of treatment (i.e. more variation due to treatments than due to error)
- Values we want to be small: p value, Sum Sq and Mean Sq of residuals (i.e. less variation due to error)

5. THE NUMBER THREE

Having more than three samples is enough to start a T-test (statistic test), why? Why less than three is not okay?



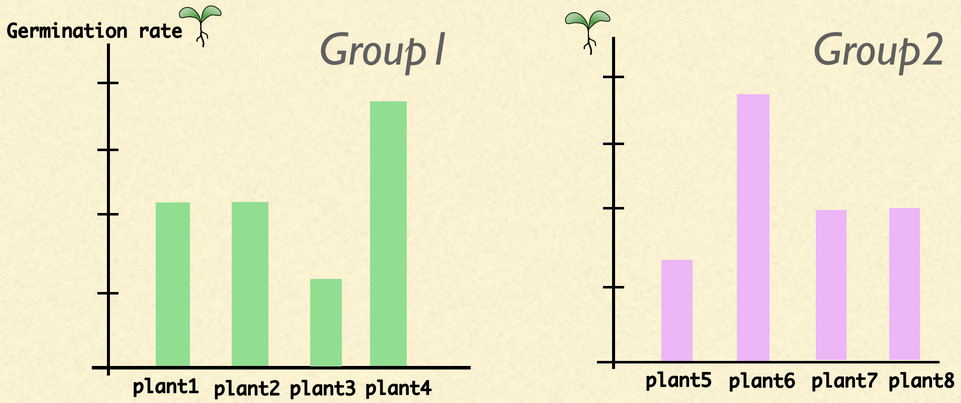
Only having at least three samples give you the possibility to draw this "normal distribution" curve.

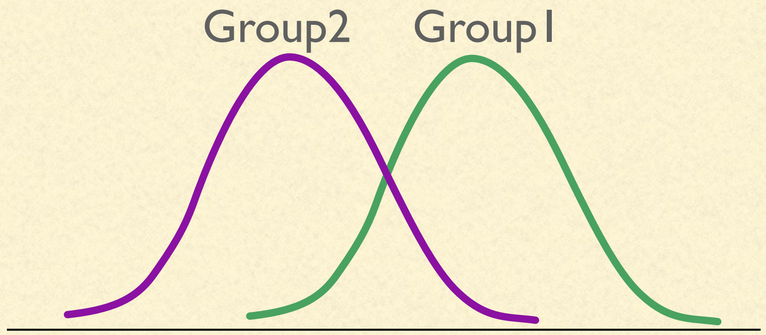


The "shape" of the curve cannot be known with just one or two samples. However, as you can see, the more samples you have, the more reliable/accurate the curve drawing is.

5. THE NUMBER THREE

And why do we need the curve?



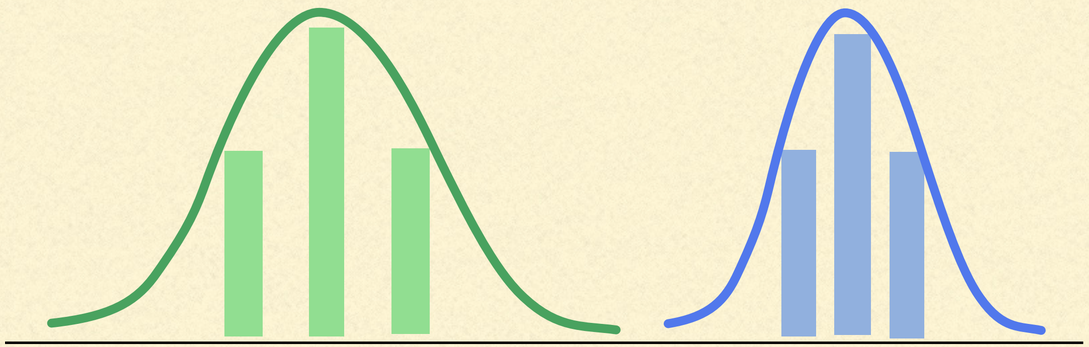


The distribution and overlapping of the curves make the statistical test result

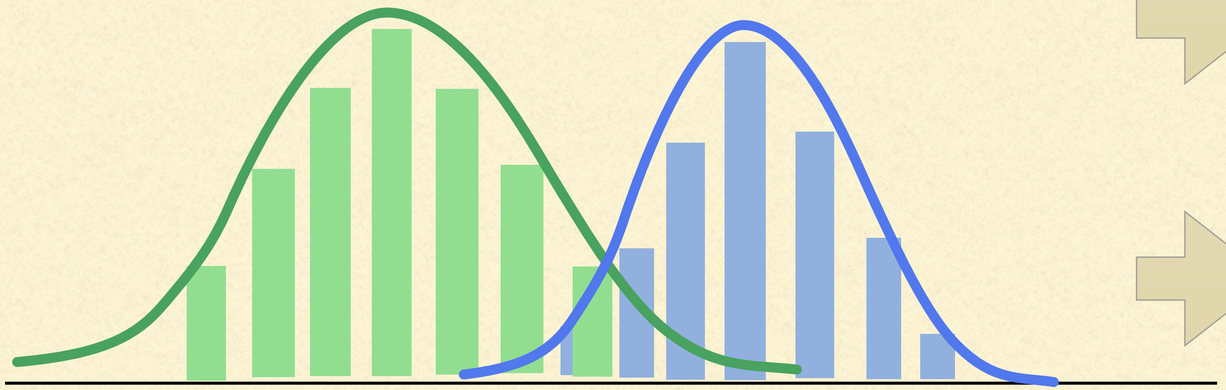
Are two groups the same? Are they different? We cannot just say they overlap a little bit… *Statistics is the grammar of science!*

5. THE NUMBER THREE

Larger sample size increases the statistical power - what does it mean?



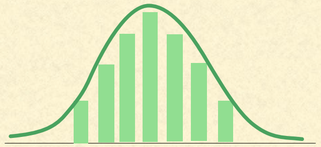
They are not overlapping at all. For sure the test says they are different.



With larger sample size, we could have seen them overlapping.

The sample size is about reliability of your result

6. VARIANCE AND STANDARD DEVIATION



For any population, you have these **parameters**:

$$Variance = \sigma^2 = \frac{\Sigma(x_i - \mu)^2}{n}$$

$i = 1, 2, 3, 4, ...$

μ : *population mean*

n = *number of datapoints*

$$Standard\ Deviation = \sqrt{Variance} = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(x_i - \mu)^2}{n}}$$

$i = 1, 2, 3, 4, ...$

μ : *population mean*

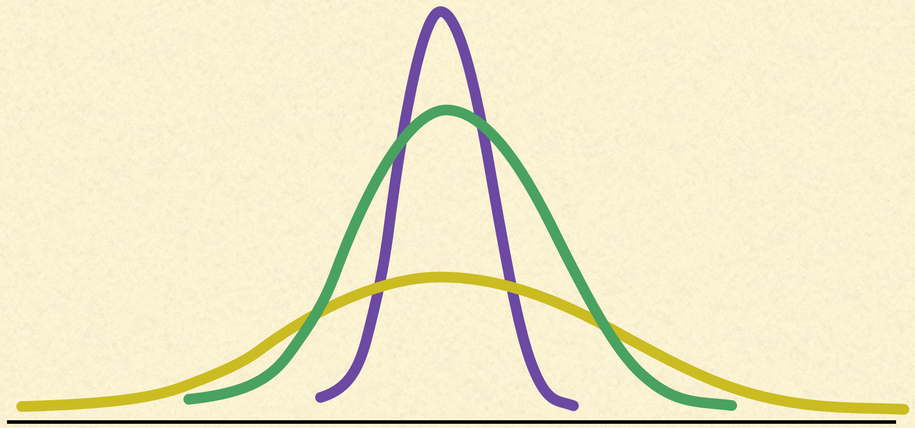
n = *number of datapoints*

6. VARIANCE AND STANDARD DEVIATION

What are the difference and why do they tell us?

- They're different and related mathematically: variance is the squared difference from the mean, standard deviation is the square root of variance

- Variance tells the dispersion of the data, Standard Deviation measures the spread of the data around the mean



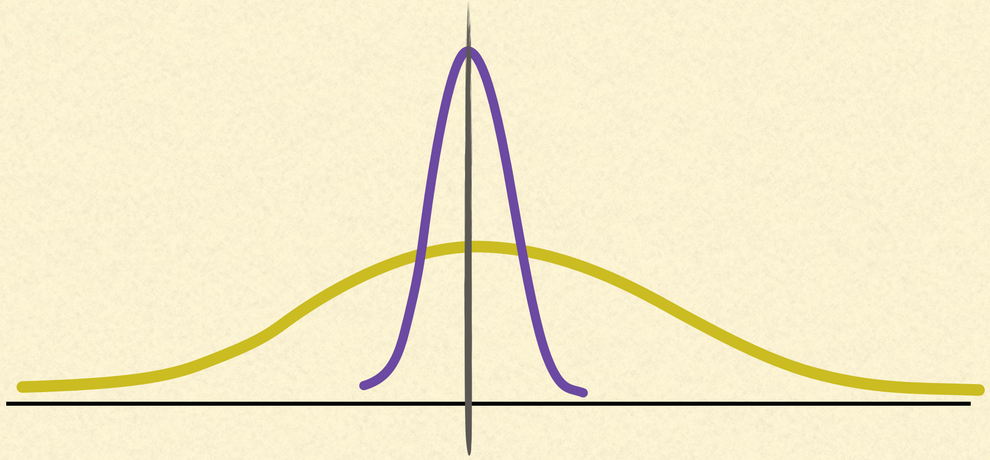
It helps us in understanding our data,
selecting the samples, thus making decision.

TEST YOURSELF

Questions

1. If two datasets have the same mean but different standard deviations, one with higher standard deviation would spread [*more* or *less*] around the centre?

2. A dataset has low standard deviation, this means the data points are [*dense near the mean* or *widespread from the mean*]?

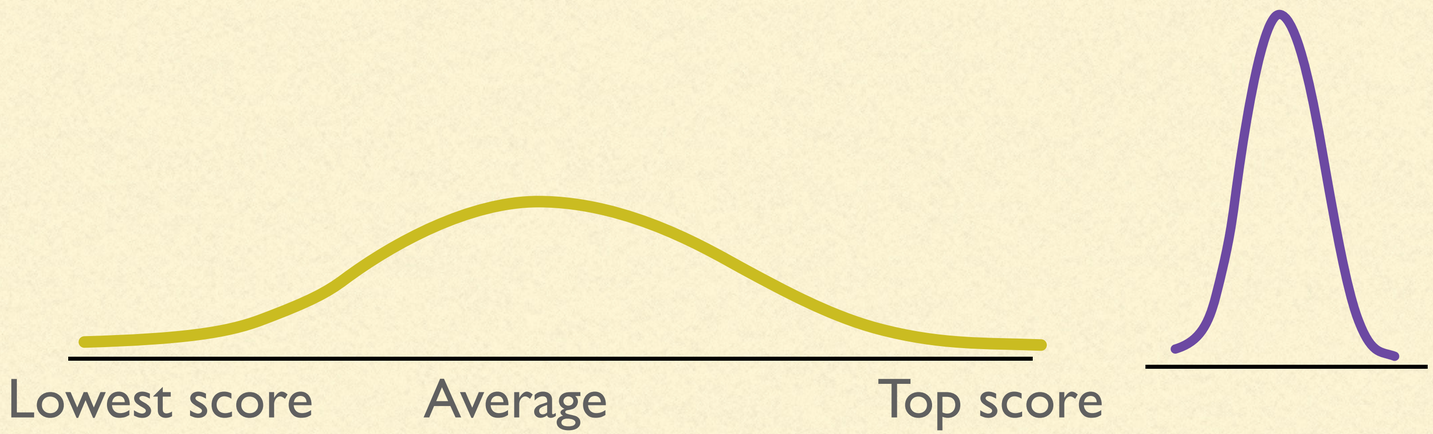


TEST YOURSELF

Questions

3. If the values in the dataset are all equal, what the values of Variance and Standard Deviation are?

4. After an exam, the grades of the students in Class 1 follow this yellow distribution, and Class 2 the purple. Describe what it means here?

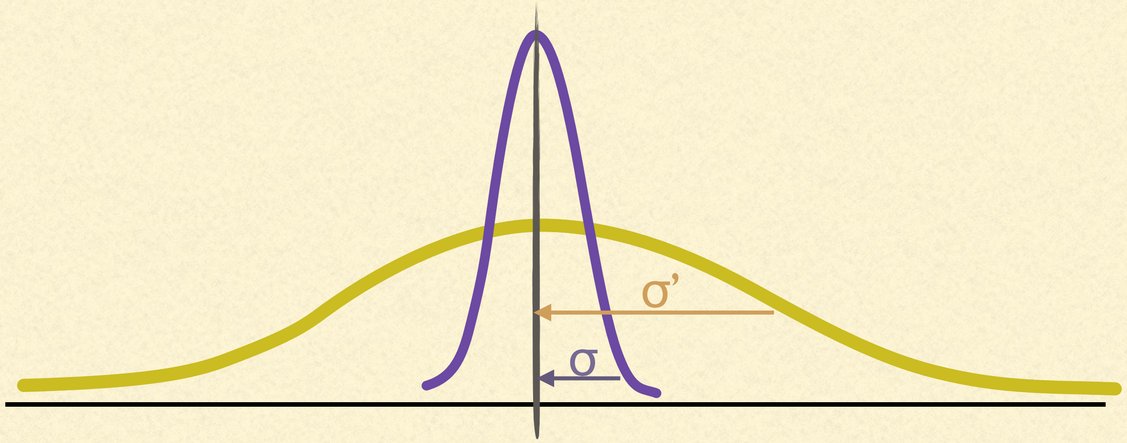


TEST YOURSELF

Answers

1. If two datasets have the same mean but different standard deviations, one with higher standard deviation would spread *more* around the centre

2. A dataset has low standard deviation, this means the data points are *dense near the mean*



TEST YOURSELF

Answers

3. If the values in the dataset are all equal, what the values of Variance and Standard Deviation are?

=> *They would be both zero.*

4. After an exam, the grades of the students in Class 1 follow this yellow distribution, and Class 2 the purple. Describe what it means here?

In class 1, the top-score and lowest students are further deviated from the mean. In this class, we find all sorts of students ranging from who perform well to who not so well. While in class 2, all scores centre around the mean, meaning in Class 2 most of the students perform at the same level.

“It’s easy to lie with statistics. It’s hard to tell the truth without statistics”

— Andrejs Dunkels