ANALYSIS OF VARIANCE

Tutorial #7

FIRST: SCIENTIFIC EXPERIMENTS

- Response variable is explained by the explanatory variable
- response ~ explanatory

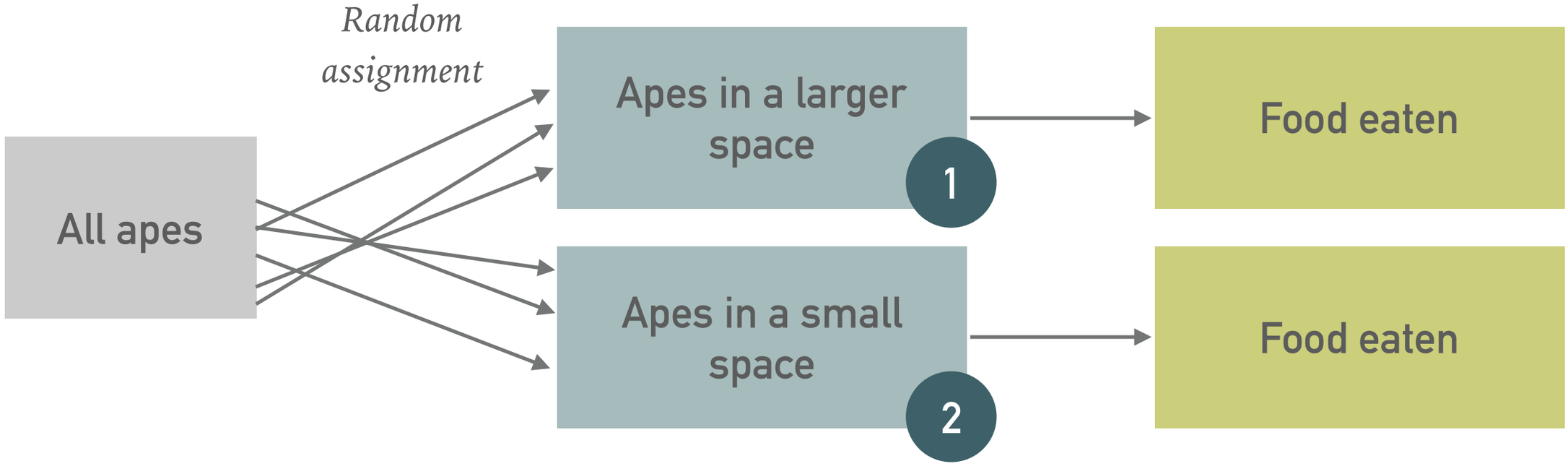
Explanatory variable

"Predictor"

Response variable

EXPERIMENTS

- **Research question:** Do apes in a smaller space eat less than apes living in a larger space?
- **Method:** two categorical predictors, so test it with a **t-test**



EXPERIMENTS: DO IT YOURSELF!

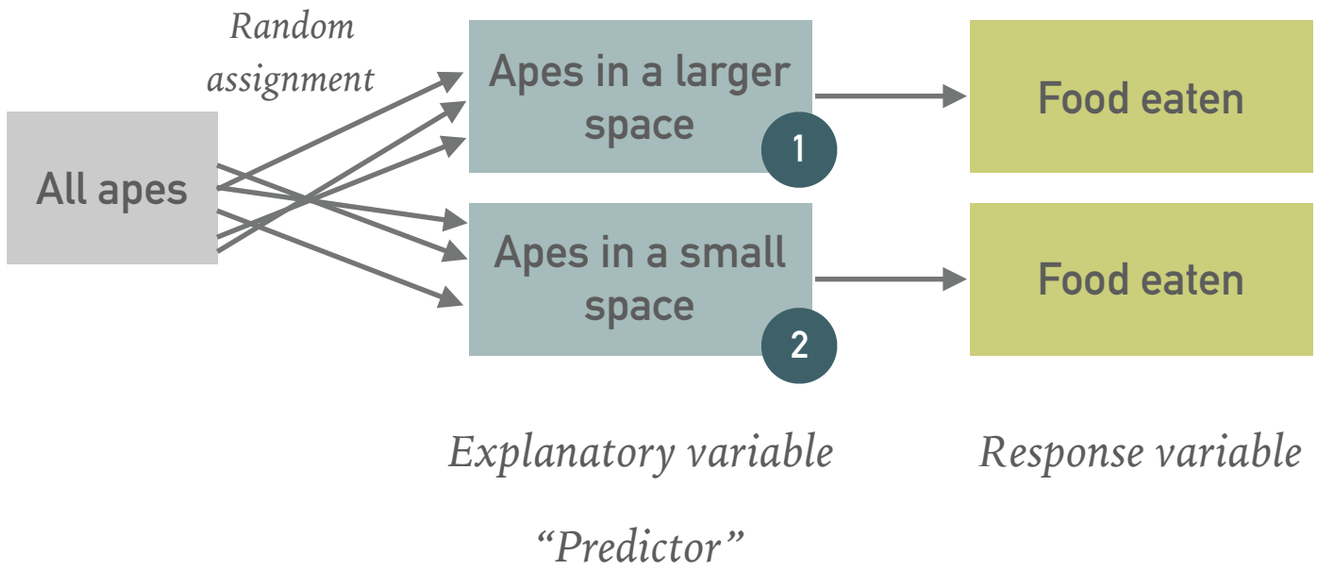
➤ Draw experimental designs for the following questions.

1. **Research question:** Where does the Hippeastrum plant grow larger, in mountains or in the valley?

2. **Research question**: Are environmental scientists happier than other students?

3. **Research question**: Do recreational parks have positive health benefits?

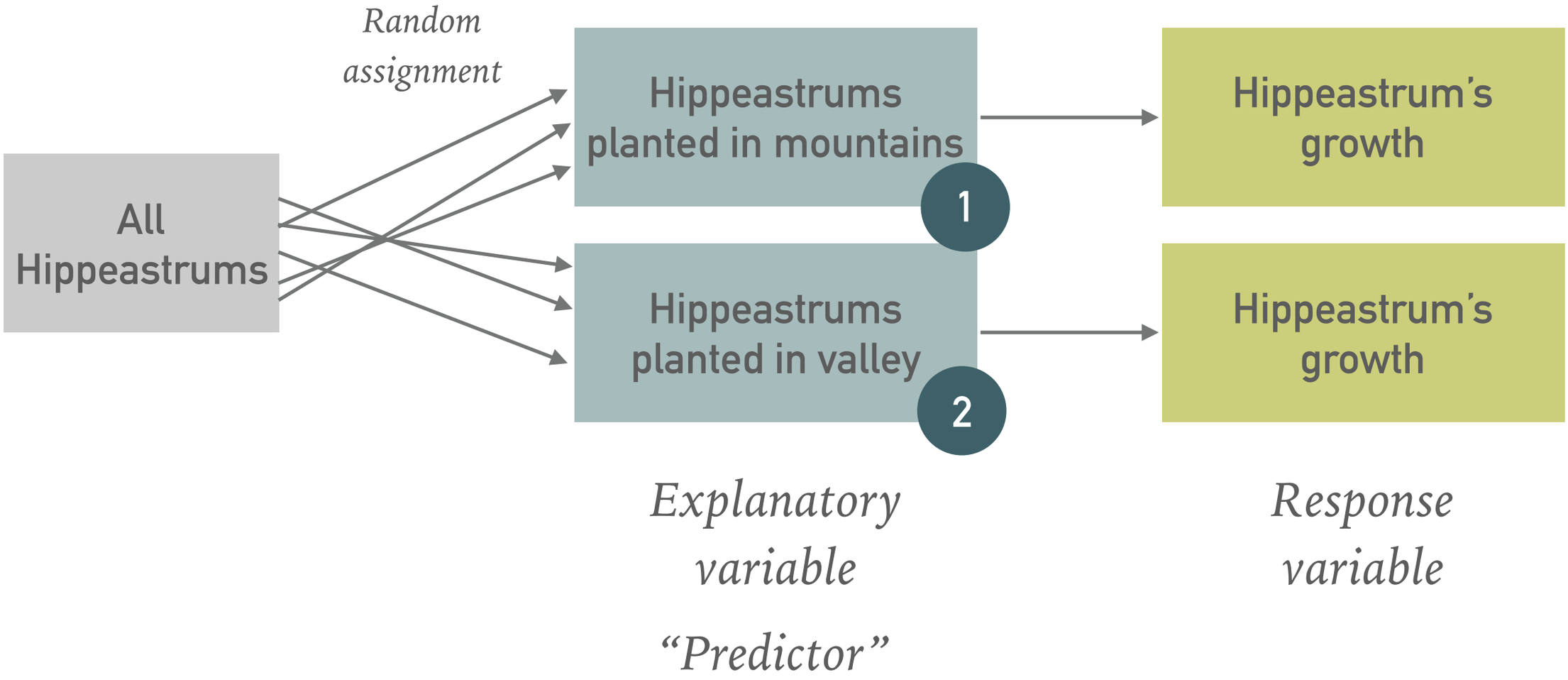
4. **Research question**: Which of the 4 fertilizers works best?



SOLUTIONS

1. **Research question:** Where does the Hippeastrum plant grow larger, in mountains or in the valley?

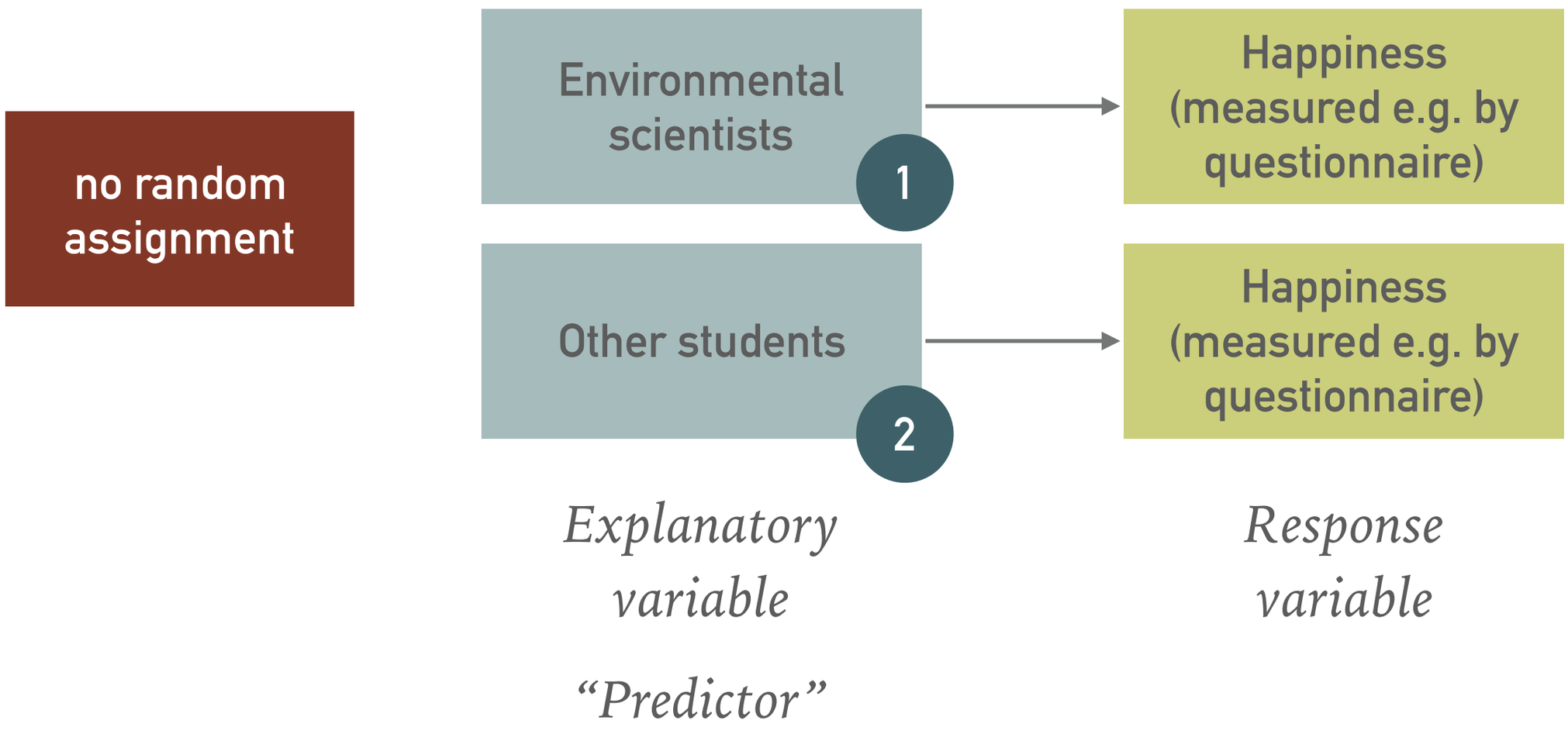
➤ **Method:** two categorical predictors, so test it with a **t-test**



SOLUTIONS

2. **Research question**: Are environmental scientists happier than other students?

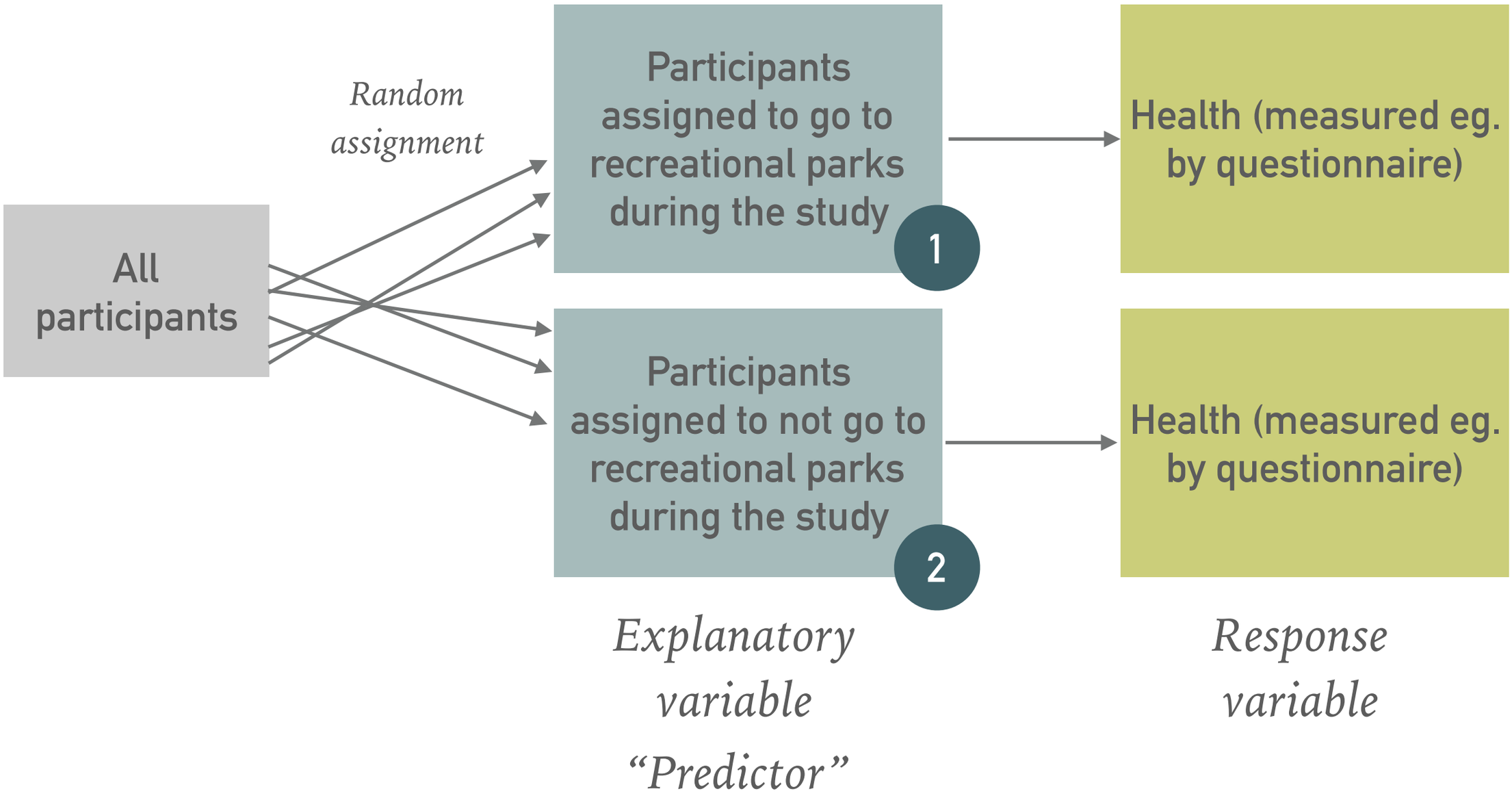
➤ **Method:** two categorical predictors, so test it with a **t-test**



SOLUTIONS

3. **Research question**: Do recreational parks have positive health benefits?

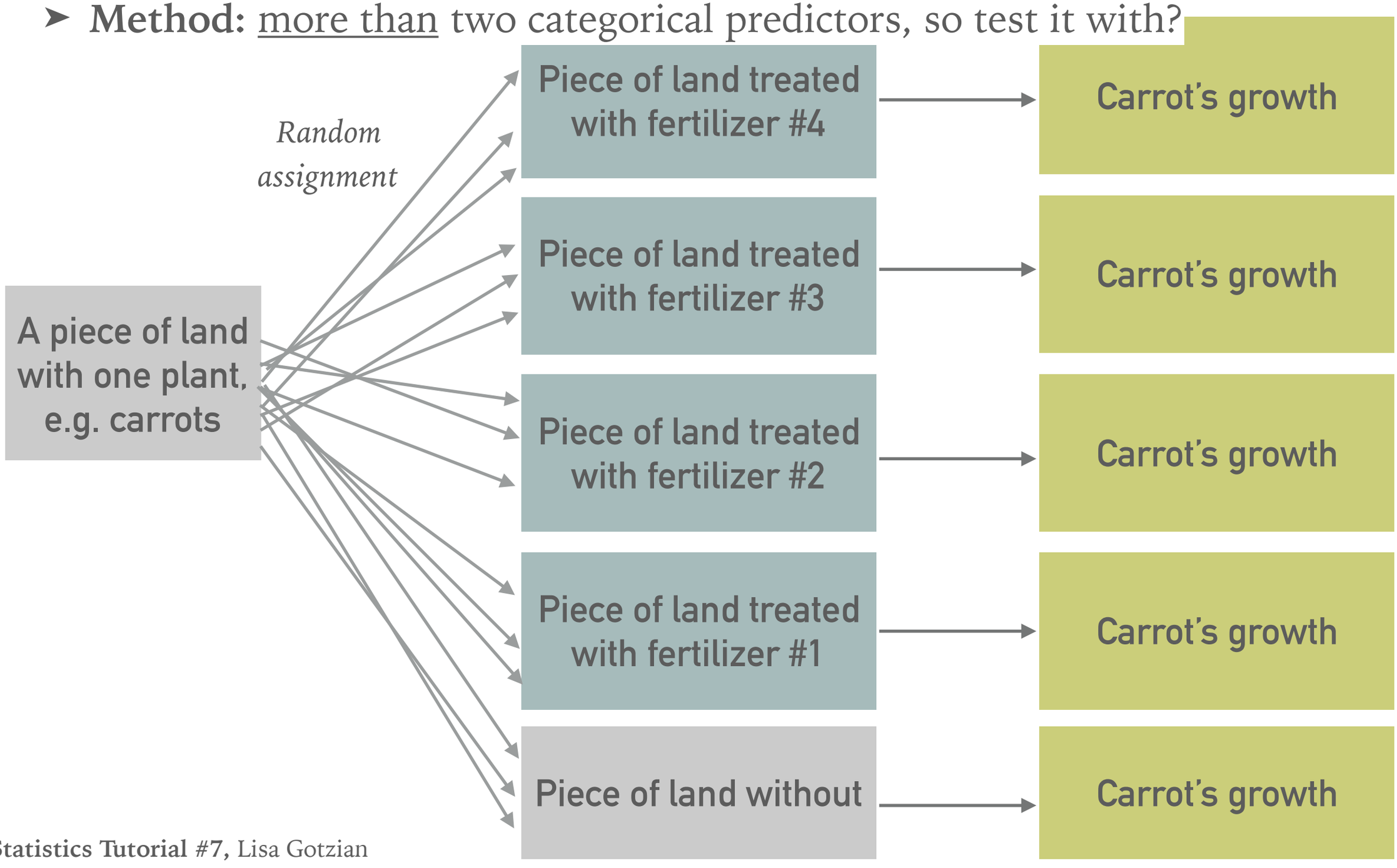
➤ **Method:** two categorical predictors, so test it with a **t-test**



SOLUTIONS

4. **Research question:** Which of the 4 fertilizers works best?

➤ **Method:** more than two categorical predictors, so test it with?



ANOVA

1. **Research question:** Which of the 4 fertilizers works best?

➤ **Method:** more than two categorical predictors, so test it with?

  ➤ several t-tests or

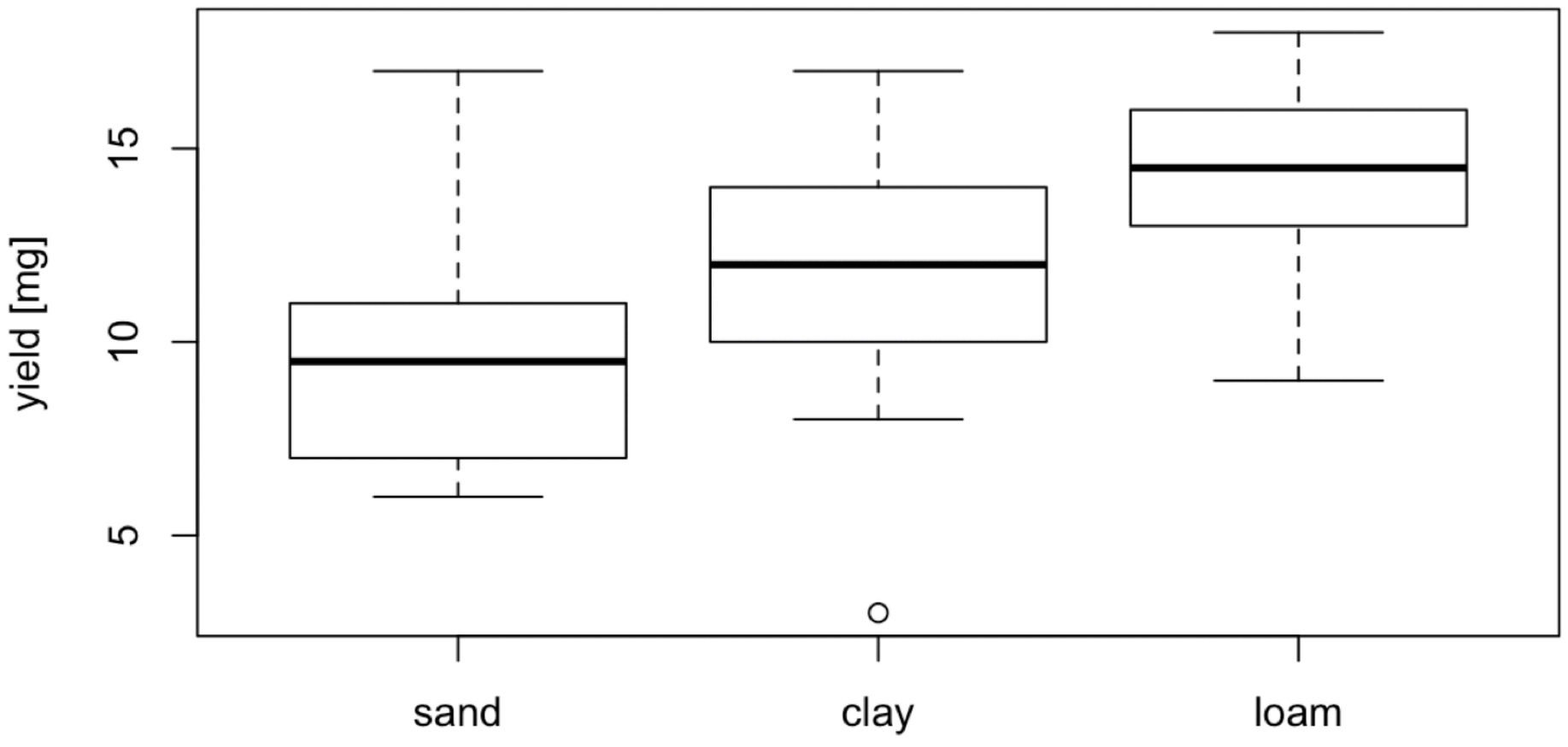
  ➤ ANOVA

Test	Data type	Purpose	Null Hypothesis	Alternative Hypothesis
t-test	Predictor: Categorical Response: Interval/Ratio	2 samples significantly **different**?	The two samples means are not different.	The two samples means are different.
ANOVA	Predictor: Categorical Response: Interval/Ratio	more than 2 samples statistically **different**?	Not different.	Different.

ANOVA

- ANOVA = Analysis of variance
- compare the means of more than two groups

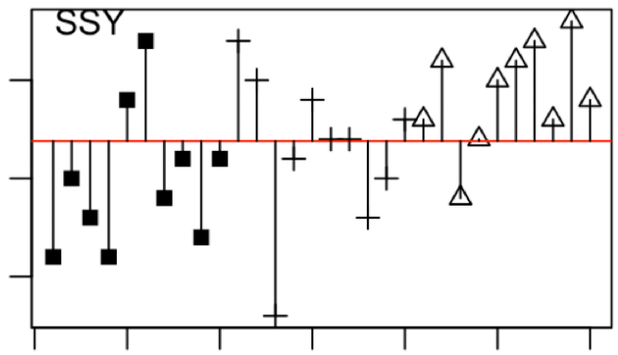
```
boxplot(yield~soil, data = yield, ylab="yield [mg]")
```



DERIVING HOW IT IS CALCULATED

- In the following slides, we'll derive how to determine that the **means are different** aka how an ANOVA is done.
- 1a) The distance between points and the mean squared is the **Sum of Squares**

$$SS = \sum (y_i - \bar{y})^2$$



- 1b) We will **split** the overall Sum of Squares

overall Sum of Squares → *treatment Sum of Squares*

overall Sum of Squares → *error Sum of Squares*

- 2) Then, we'll calculate the **ratio** of the two…
- 3) that will give us the **p-value.**

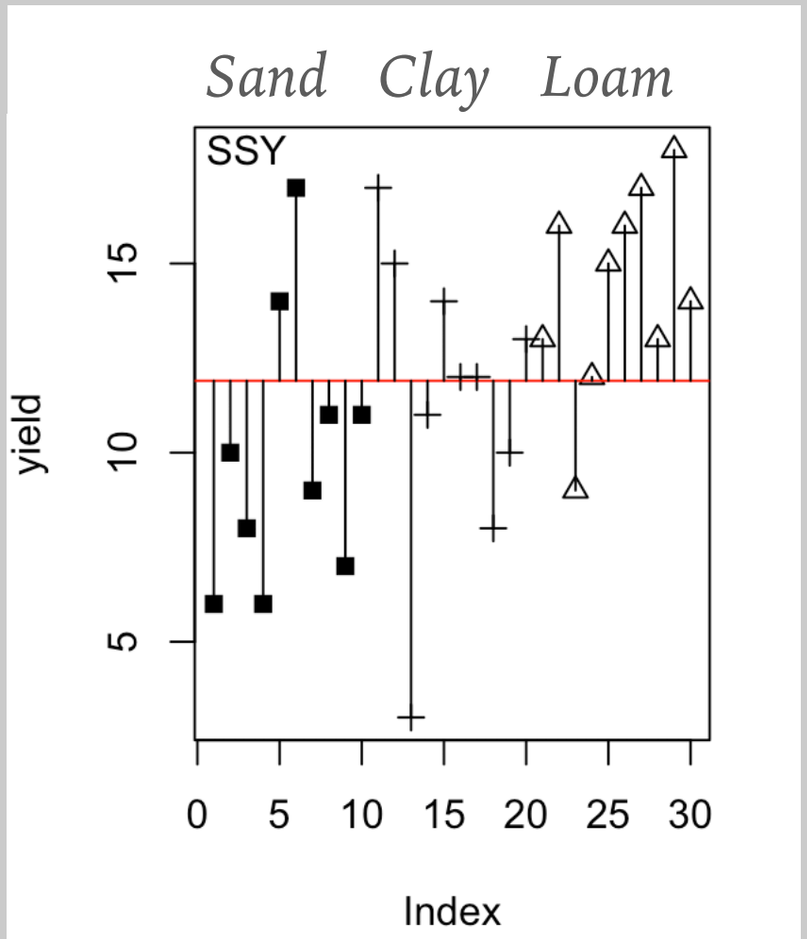
THE SUM OF SQUARES

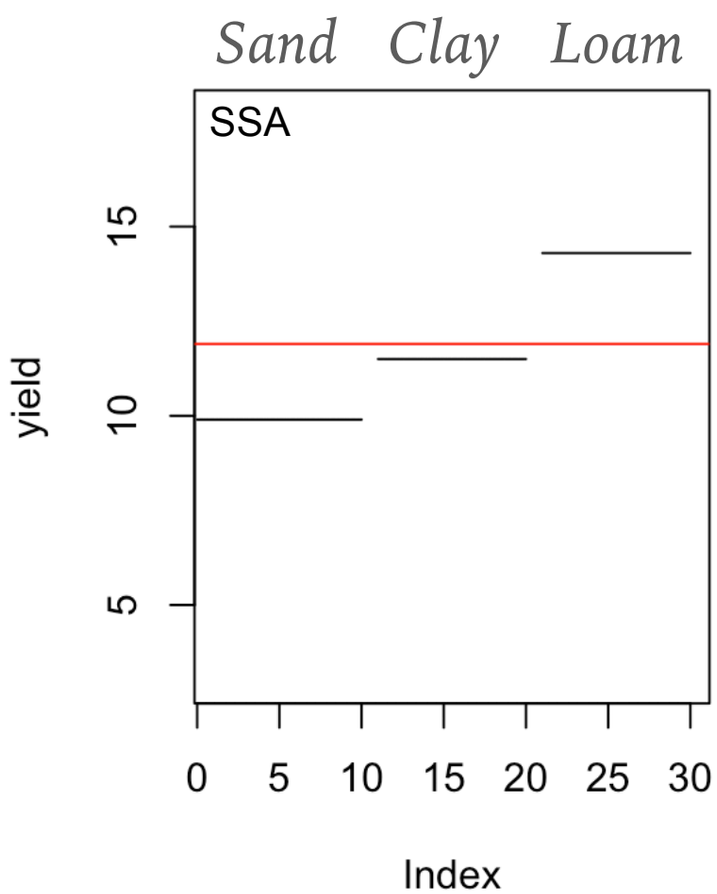
total SS = ***treatment SS*** + ***error SS***

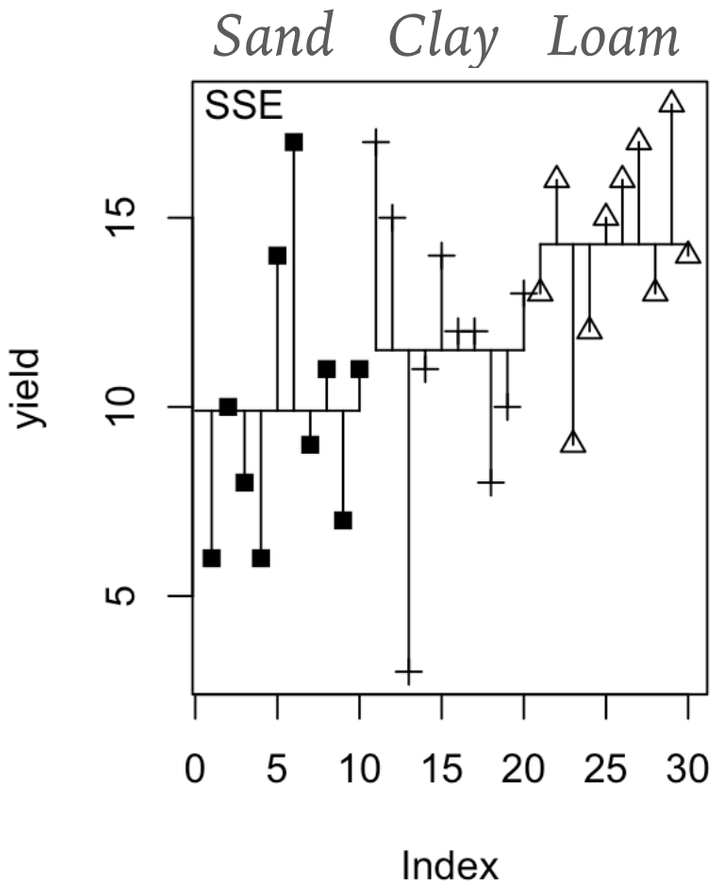
difference data point - overall mean

difference treatment mean - overall mean

difference data point - treatment mean







FROM THIS, WE INFER THE TEST STATISTIC

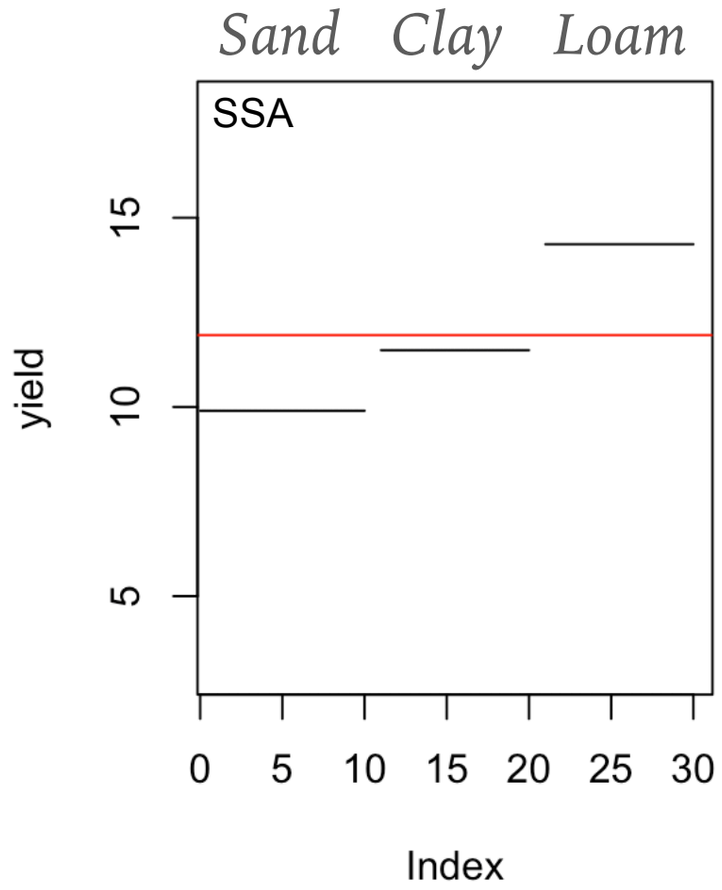
2

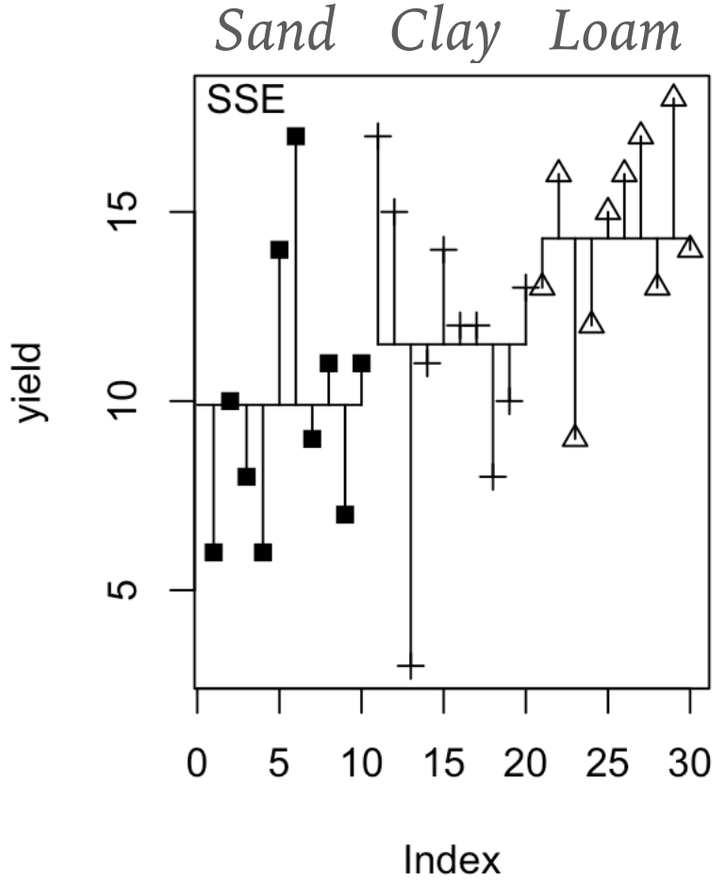
$$\textit{F ratio} = \frac{\textit{treatment SS}}{\textit{error SS}}$$

test statistic

difference treatment mean - overall mean

difference data point - treatment mean





THE P-VALUE

$$\textit{F ratio} = \textit{treatment SS} / \textit{error SS}$$

test statistic

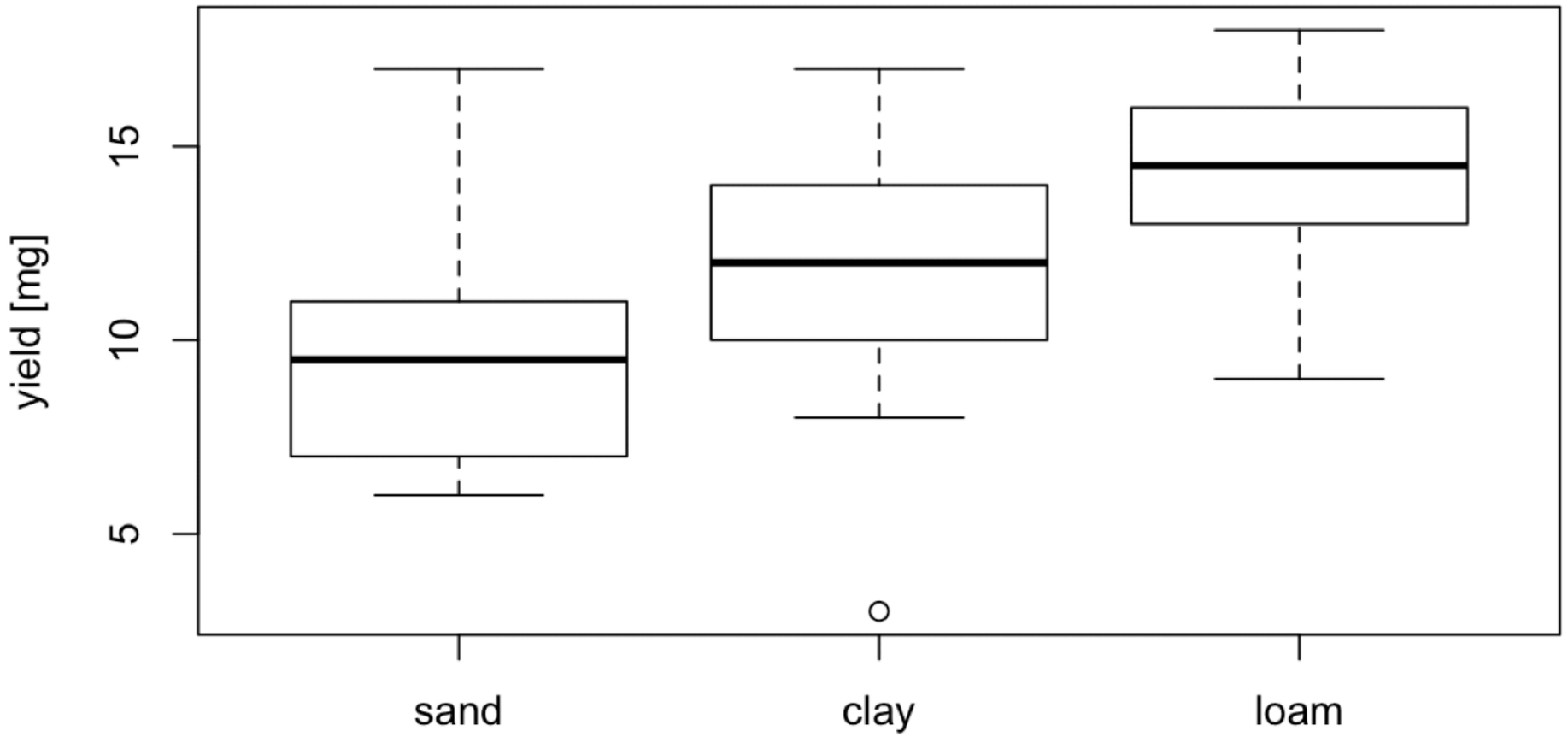
```
> anova(model)
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value  Pr(>F)
soil       2   99.2  49.600  4.2447 0.02495 *
Residuals 27  315.5  11.685
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

H0: No difference in means.

H1: There is a difference in means.

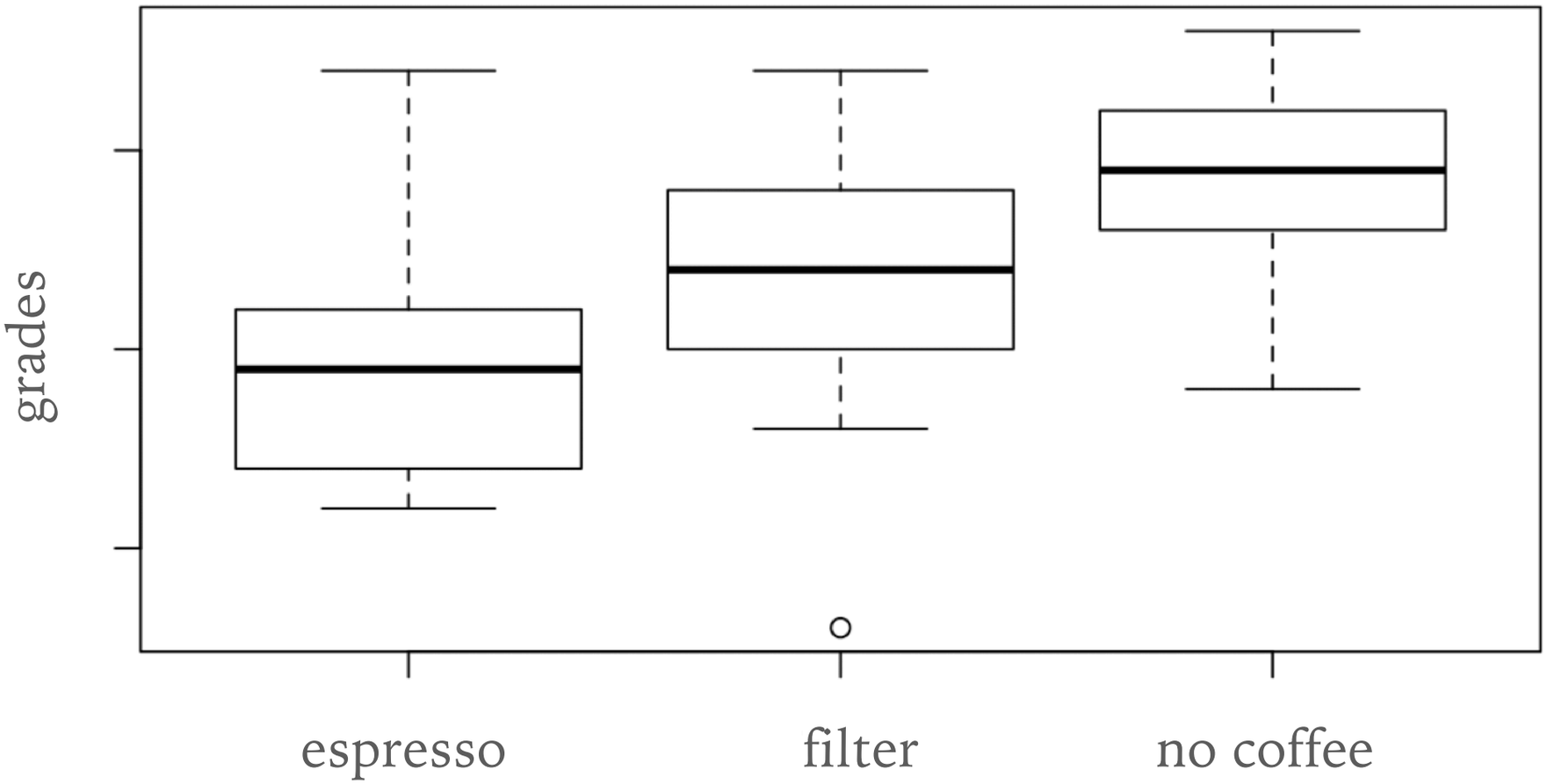
BOXPLOT OF THE 3 TREATMENTS

$p < 0.05 = The\ means\ are\ significantly\ different.$



COFFEE EXAMPLE

$p < 0.05$ = *The means are significantly different.*



imagine this coffee example

THERE'S DIFFERENT VERSIONS OF ANOVAS

- [**t-test**: one factor with two levels]
- **One-Way Anova**: one factor with three or more levels
- **Two/Three-Way Anova**: two or three factors with levels
- **Factorial design**: <u>replicate</u> the levels

Coffee/no coffee

Espresso/Filter/No coffee

Espresso/Filter/No coffee
for the second semester people

Espresso/Filter/No coffee
for the fourth semester people

THE RELEVANT R COMMANDS

```
model <- lm(yield ~ soil, data = yield)
anova(model)
```

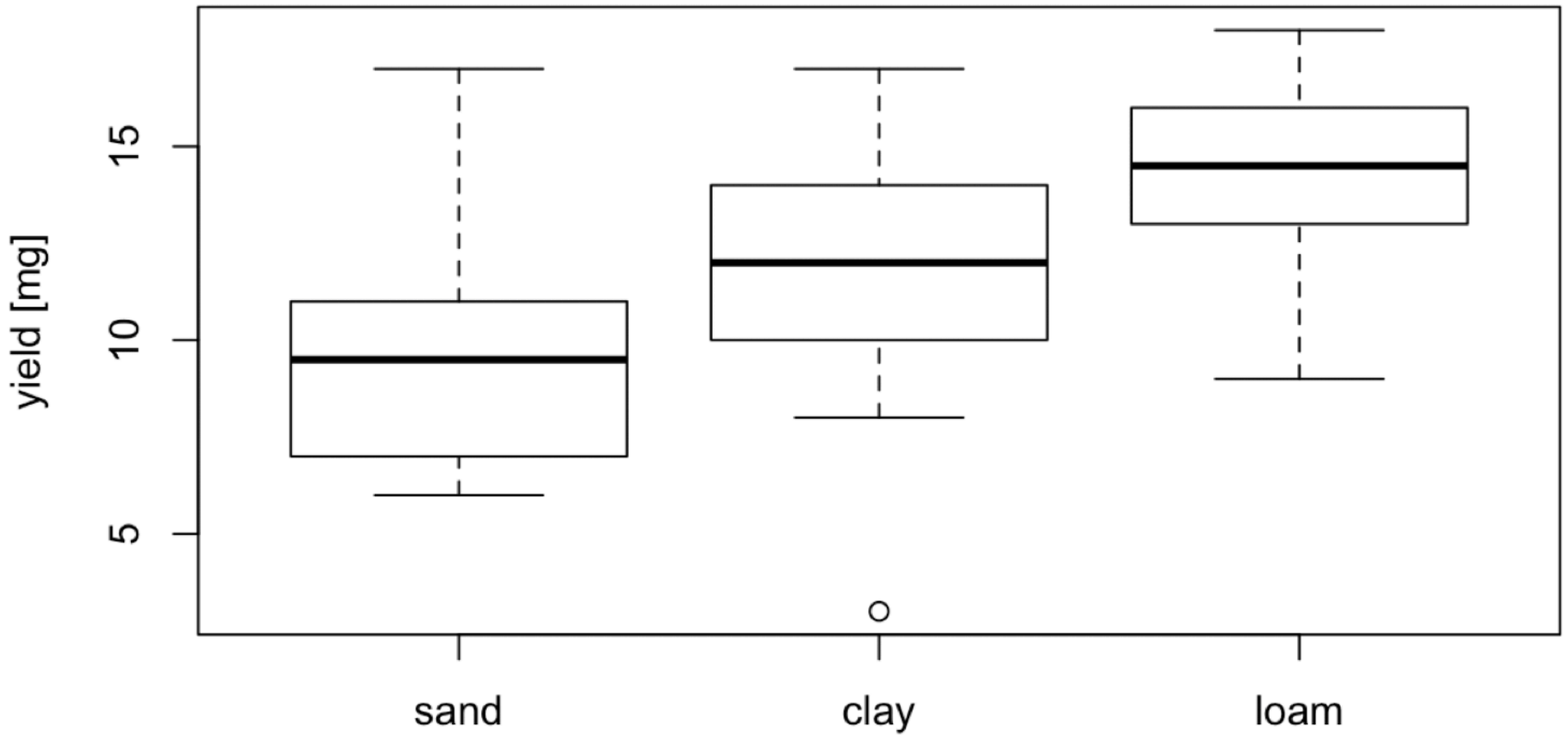
option #1

```
model <- aov(yield ~ soil, data = yield)
summary(model)
```

option #2

BOXPLOT OF THE 3 TREATMENTS

$p < 0.05 = The\ means\ are\ significantly\ different.$



WHICH MEANS ARE DIFFERENT?

- Now that you know that the means are different, you need to know *which* means are different.
- For a pair-wise comparison, you need Post-Hoc Tests, in this case the *TukeyHSD*.

pairwise comparisons

```
> TukeyHSD(model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = yield ~ soil)

$soil
           diff        lwr      upr     p adj
clay-sand   1.6 -2.1903777 5.390378 0.5546301
loam-sand   4.4  0.6096223 8.190378 0.0204414
loam-clay   2.8 -0.9903777 6.590378 0.1785489
```

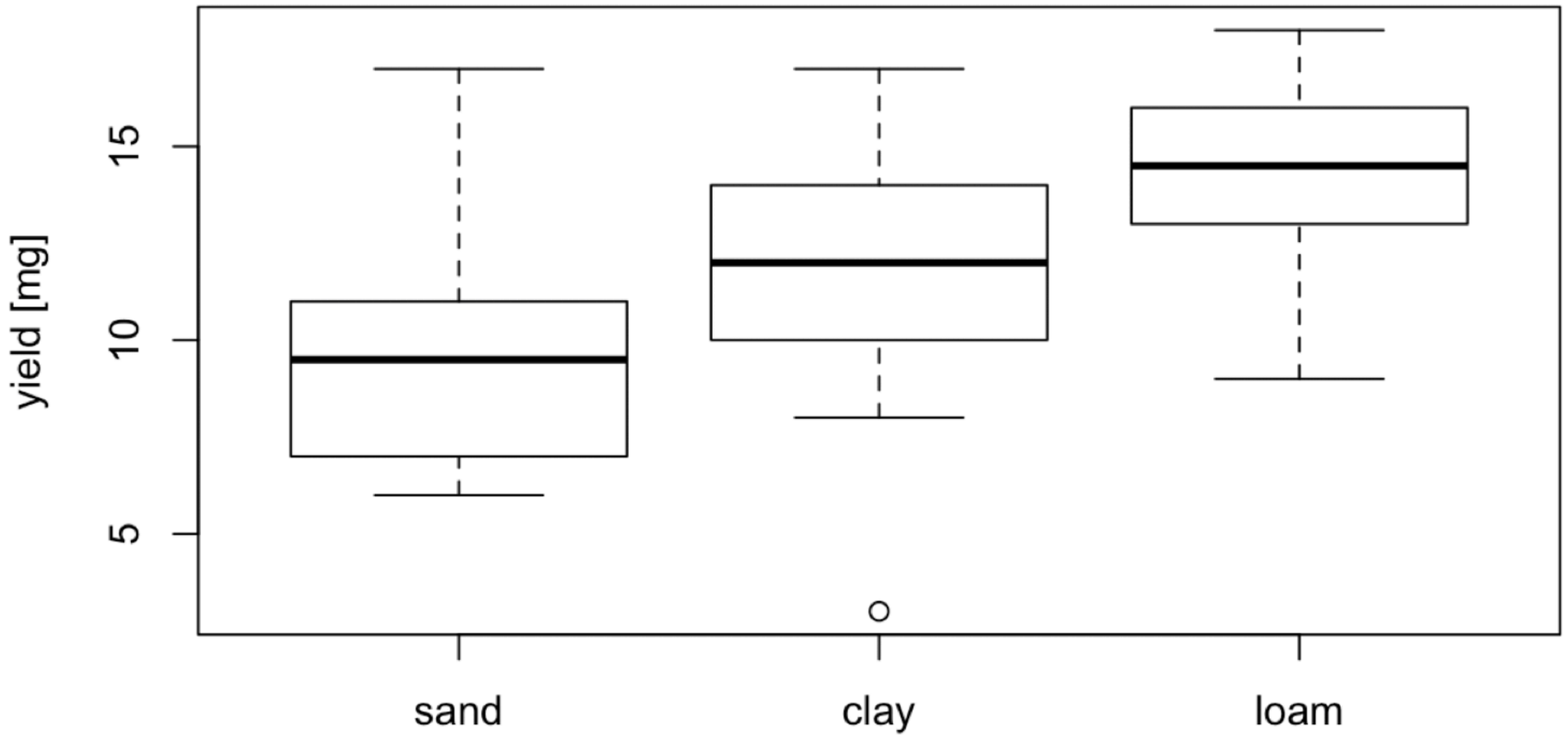
THE PROCESS OF AN ANOVA

Before	test if all variables are normally distributed	Shapiro-Wilk Test or Kolmogorov-Smirnov Test (does the same, just a new name ;))
	test if variances are homogenous	Levene's test/Brown–Forsythe test (F test only works for two groups)
	test if means are different	ANOVA
After	test which means are different	Post-Hoc-Test -> Tukey HSD

DONE WITH THEORY

PRACTICE DRAWING A BOXPLOT

- https://www.youtube.com/watch?v=09Cx7xuIXig - Khan Academy on drawing boxplots



Next time: **Scientific papers**

Which is in two weeks - see you there!