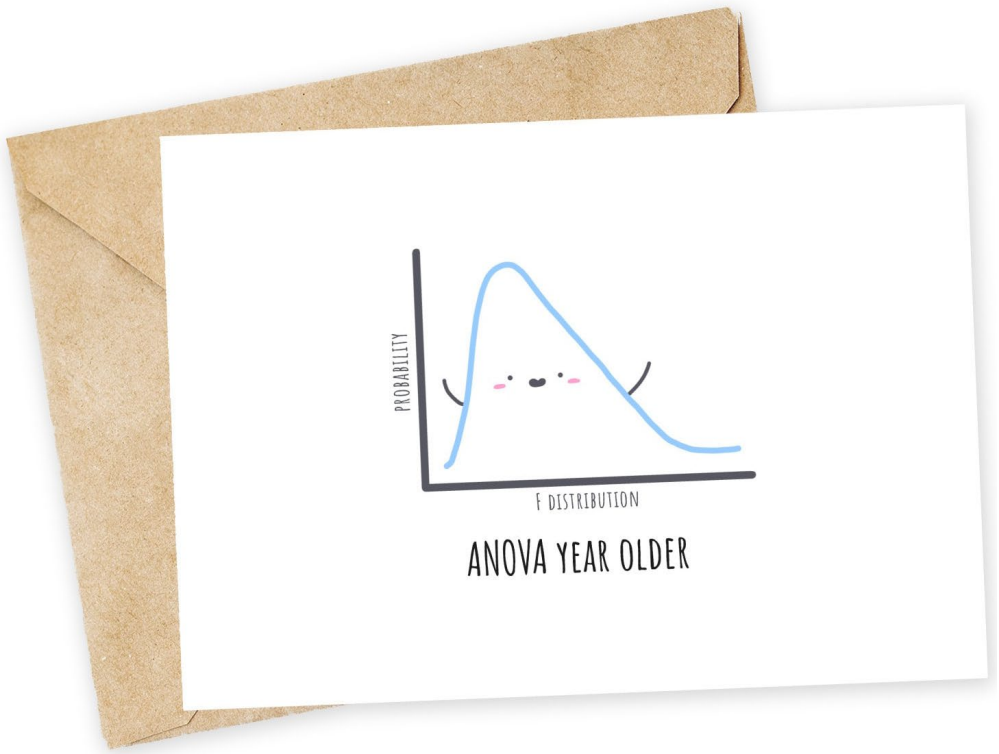
Comparing Three or More Groups

ENTMLGY 6707 Entomological Techniques and Data Analysis



Questions to ponder

- What is one kind of analysis you are going to conduct in graduate school?
- How did you decide it was the right one?
- Are there other options to analyze the data? If so, how will you justify your decision?

Learning objectives

1. Understand when and how to conduct an ANOVA
2. Summarize experimental components using linear models

Which analytical framework should I use?

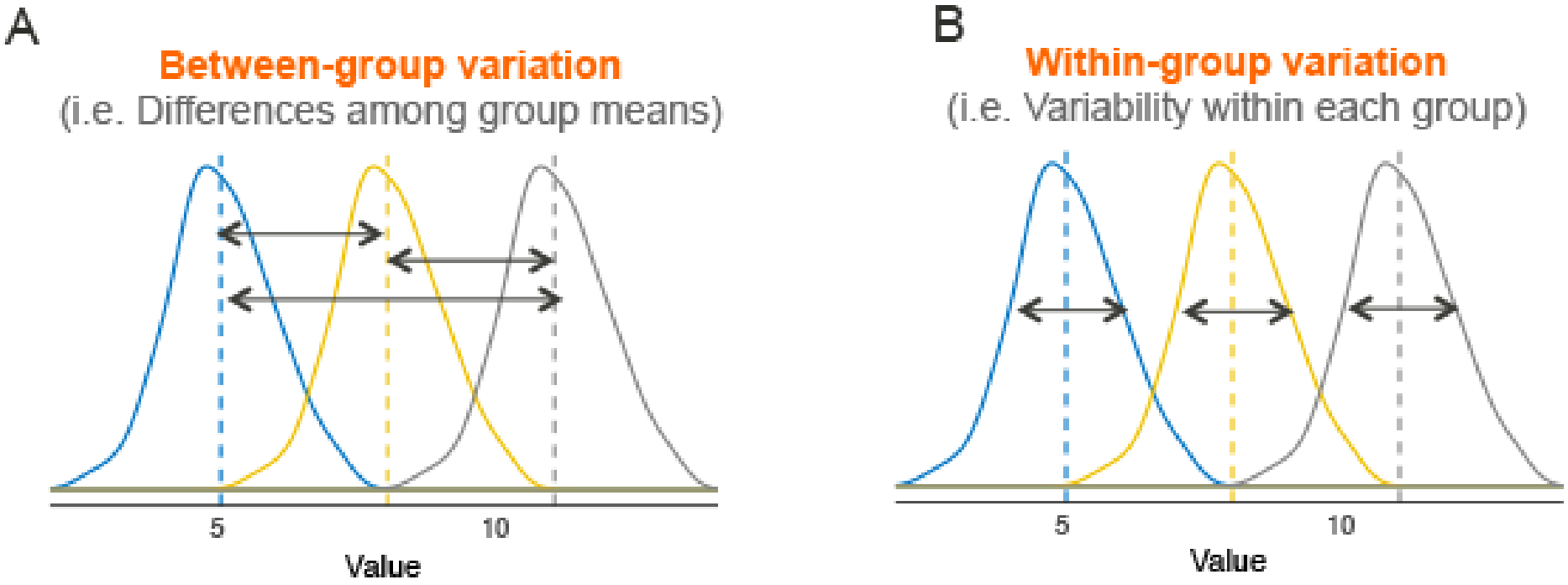
Table 4.1 Examples of the Generalized Linear Model as a Function of Independent Variab

Predictors (X)	Responses (Y): *Continuous DV*	*Binary DV*	*Unordered Multicategory DV*	*Ordered Categorical DV*	*Count DV*
Continuous IV	OLS regression	Binary logistic regression	Multinomial logistic regression	Ordinal logistic regression	OLS, Poisson regression
Mixed continuous and categorical IV					
Binary/ categorical IV only	ANOVA and *t*-test	Log-linear models	Log-linear models		Log-linear models

ANOVA, analysis of variance; DV, dependent variable; IV, independent variable; OLS, ordinary least squares.

Chapter 4: Simple Linear Models With Continuous Dependent Variables: Simple ANOVA Analyses
In: [Regression & Linear Modeling: Best Practices and Modern Methods]

ANalysis **O**f **VA**riance (ANOVA)



Writing tip: use "between" when referring to two groups, use "among" when referring to three or more groups

ANalysis **O**f **VA**riance (ANOVA)

Source	df	SS	MS	*F*
Treatment	*k*-1	formula	SSTrt1÷dfTrt1	MS_{trt}/MS_{err}
Error	*N*-*k*	formula	SSErr÷dfErr	
Total	*N*-1	formula	-	

Note that:
MS = mean square = sum of squares ÷ degrees of freedom for each component.

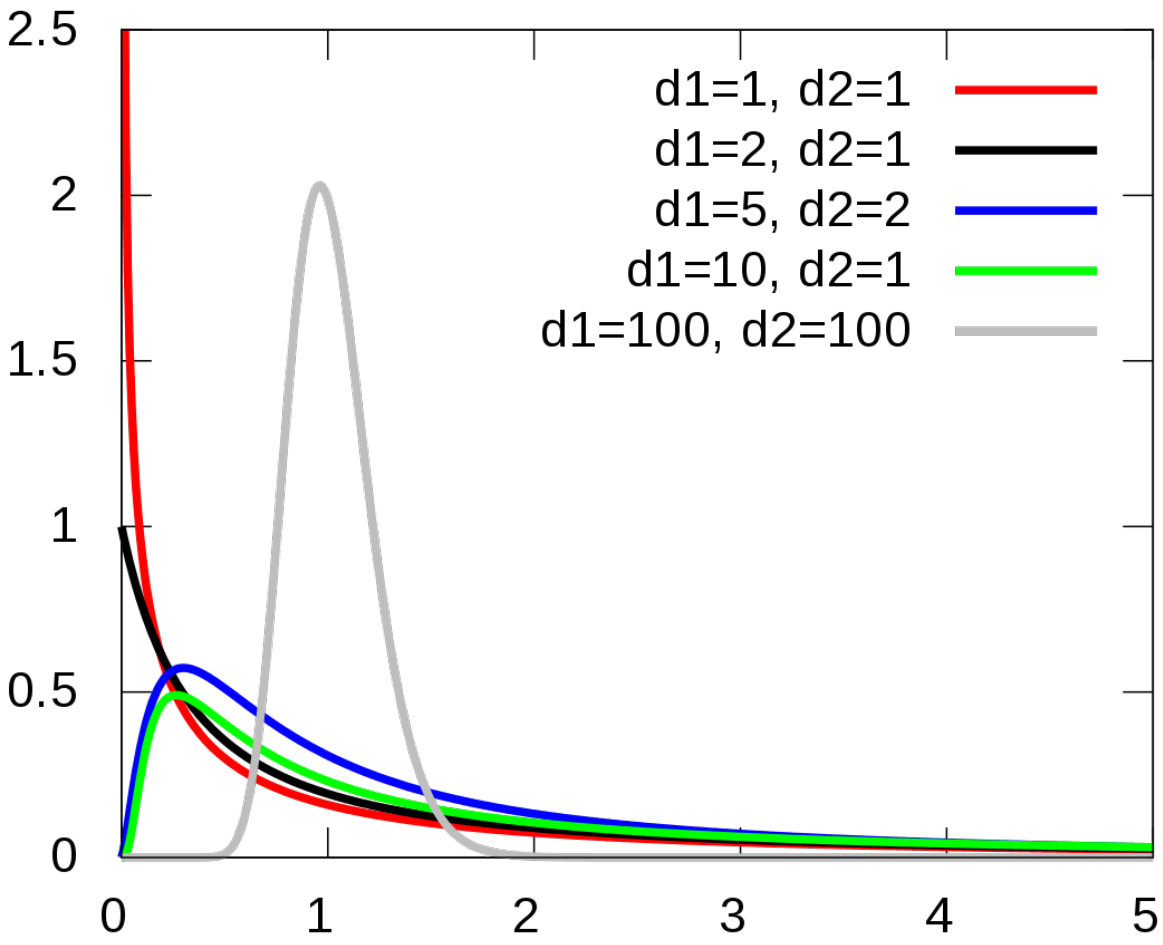
We get *F*-statistics from this table (*F*-ratio)!

$F_{\text{trt df, err df}} = MS_{trt} / MS_{err}$

From the *F*-statistic and degrees of freedom, we can calculate a *p*-value and evaluate our null hypothesis.

F distribution

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} \; d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x \, \mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$



A quick review

$\hat{Y}$ or $\hat{X}$ pronounced "Y-hat" or "X-hat" indicates an estimate. That is, $\hat{Y}$ is our "best guess" for the true value of Y.

Let's say we sample from a population of ponderosa pine trees and get the following diameters (cm) measured at breast height (DBH; measured 1.5m above ground):

9, 17, 11, 18, 19

Our expected value, E(X), of a random draw from that string of numbers = the average = 12.4 cm

So, $\bar{X} = \hat{\mu} = \text{estimate of true population mean} = 12.4 \text{ cm}$

Linear model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Table 5.1. Three variance components.

	Notation	Variance in	Sum of squared deviations of	Formula
SS_{total}	SS_{total}	Y	Observed data from the mean	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
$SS_{treatment}$	$SS_{regression}$	Y explained by X	Fitted values from the mean value	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
SS_{error}	$SS_{residual}$	Y not explained by X	Observed values from fitted values	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

$\hat{Y}_i$ = the mean of treatment group i

$$SS_{total} = SS_{treatment} + SS_{error}$$

Linear model

$$Y_{ij} = \underbrace{\mu + \tau_i} + \varepsilon_{ij}$$

$$\hat{Y}_i = \mu + \tau_i$$

$$\hat{Y}_i = \mu + \tau_i$$

$$\sum (Y_{ij} - \bar{Y})^2$$

SS Total

$$\sum (\hat{Y}_i - \bar{Y})^2$$

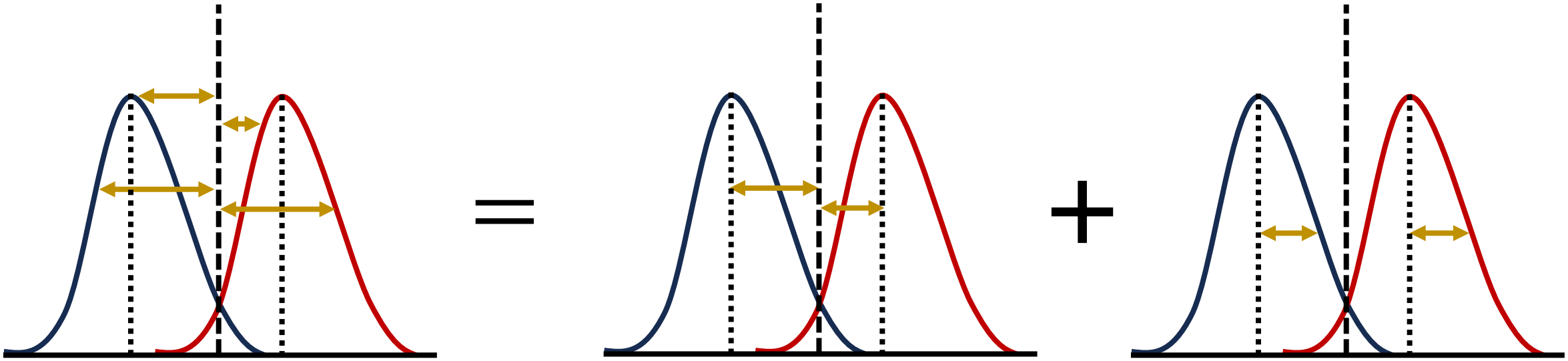
SS Treatment

$$\sum (Y_{ij} - \hat{Y}_i)^2$$

SS Error

$(\varepsilon_{ij})^2$

ANOVA



Total sum of squares = Treatment sum of squares + Error sum of squares

Sum of squares (SS) activity (don't submit)

Consider two independent samples:

A: 4, 12, 8

B: 17, 8, 11

The summary statistics are:

Mean	Variance	Sum of squares
$\bar{a}$=8	s_a^2=16	
$\bar{b}$=12	s_b^2=21	

$$\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Sum of squares (SS): Definition

A main goal of experimental design, particularly blocking, is to partition the variance into components.

Key definition:
SS Total = SS Treatment + SS Error

So, we can attribute all the variation (total sum of squares, SST) in our variable of interest (response variable) to explanatory variables (sum of squares of treatments, SST) and unexplained/leftover variation (sum of squares of error, SSE).

Sometimes written:
Total (TSS) = Explained (ESS) + Residual (RSS)

Total Sum of Squares

First, calculate mean:

$$\frac{(4 + 12 + 8) + (17 + 8 + 11)}{6} = 10$$

And SS Total is:
$(4-10)^2 + (12-10)^2 + (8-10)^2 + (17-10)^2 + (8-10)^2 + (11-10)^2 = 98$

on total df = n-1 = 5

where n = total number of observations

Treatment Sum of Squares

Treatment SS: How much of the total SS can be attributed to the differences between the two treatment groups? Replace each observation by its group mean.

A: 8, 8, 8

B: 12, 12, 12

The overall mean here is

$$\frac{(8+8+8)+(12+12+12)}{6} = 10$$

and the SS Treatment is

$(8-10)^2 + (8-10)^2 + (8-10)^2 + (12-10)^2 + (12-10)^2 + (12-10)^2 = 24$

on treatment df = number of levels of treatment – 1 = 1.

Error Sum of Squares

Error SS: How much of the total SS can be attributed to the differences within each treatment group? The SS Error is

$$(4-8)^2 + (12-8)^2 + (8-8)^2 + (17-12)^2 + (8-12)^2 + (11-12)^2 = 74$$

on error df = df Total - df Treatment = 5 - 1 = 4

ANOVA

Source	df	SS	MS
Treatment	1	24	24
Error	4	74	18.5
Total	5	98	-

Note that:
MS = mean square = SS ÷ df for each component.

We get *F*-statistics from this table (*F*-ratio)!

$F_{1,4}$=24/18.5 = 1.30

From the *F*-statistic and degrees of freedom, we can calculate a *p*-value and evaluate our null hypothesis.

Treatment Sum of Squares

Let's revisit our example when thinking about treatment SS:

A: 8, 8, 8 (from original values 4, 12, 8)

B: 12, 12, 12 (from original values 17,8,11)

Change B by **adding 6**:

A: 8, 8, 8 (from original values 4, 12, 8)

B: 18, 18, 18 (from **new** values 23,14,17)

The overall mean becomes

$$\frac{(8+8+8)+(18+18+18)}{6} = 13$$

the SS Total is now

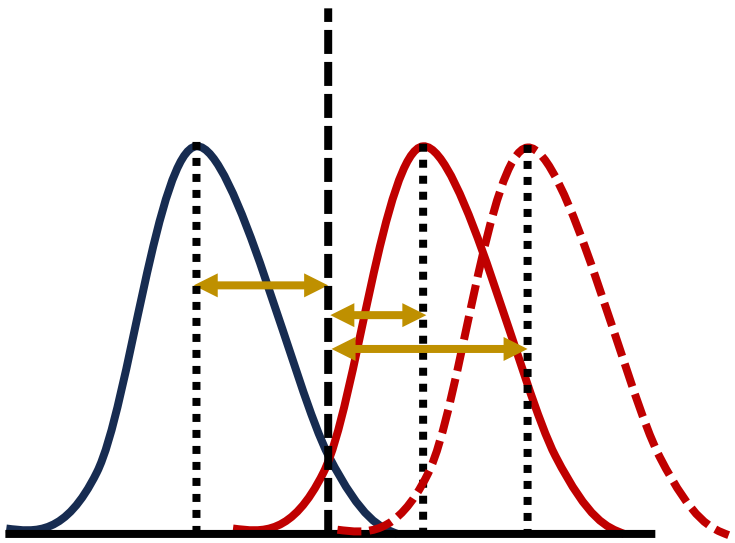
$(4-13)^2 + (12-13)^2 + (8-13)^2 + (23-13)^2 + (14-13)^2 + (17-13)^2 = 224$

the SS Treatment is now

$(8-13)^2 + (8-13)^2 + (8-13)^2 + (18-13)^2 + (18-13)^2 + (18-13)^2 = 150$

the SS Error and degrees of freedom are unchanged

$(4-8)^2 + (12-8)^2 + (8-8)^2 + (23-18)^2 + (14-18)^2 + (17-18)^2 = 74$



Adding six to each observation in B "pushed" the treatment mean for group B farther from group A, but it did not change the variance within treatment B

ANOVA

Source	df	SS	MS
Treatment	1	24 to **150**	24 to **150**
Error	4	**74**	**18.5**
Total	5	98 to **224**	-

New *F*-ratio:
$F_{1,4}$=**150**/18.5 = **8.11**

The variance between our treatments went up (i.e., the treatment means are very different), so the numerator in our *F*-ratio went up.

One-way ANOVA table

Source	df	SS	MS	F
Treatment	k-1	formula	SSTrt1÷dfTrt1	MS_{trt}/MS_{err}
Error	N-k	formula	SSErr÷dfErr	
Total	N-1	formula	-	

Two-way ANOVA tables

Table 1

Source	df	SS	MS	*F*
Treatment A	k-1	formula	SSTrt1/dfTrt1	MS_A/MS_e
Treatment B	b-1	formula	SSTrt2/dfTrt2	MS_B/MS_e
Error	N-dfTrt1-dfTrt2-1	formula	SSErr/dfErr	
Total	N-1	formula	-	

Table 2

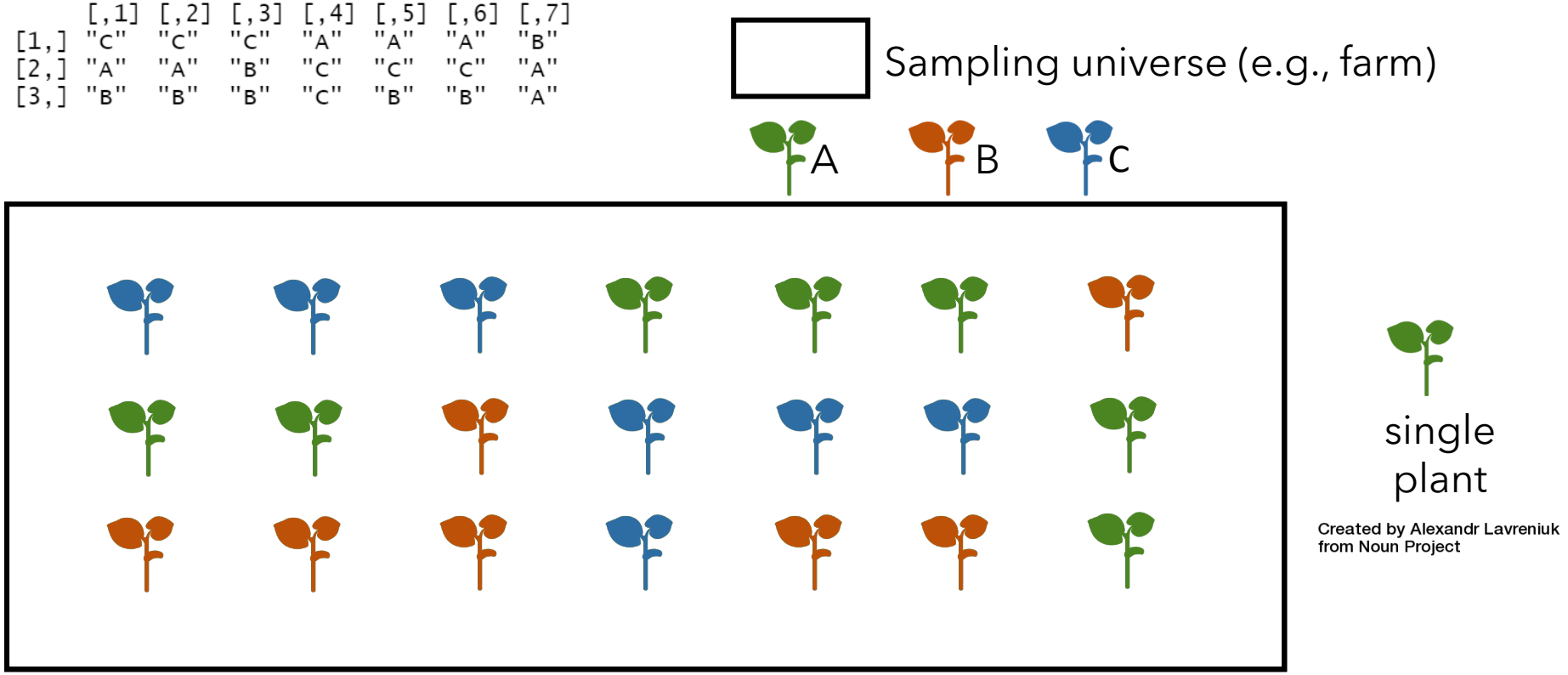
Source	Df	SS	MS	*F*
Treatment A	k-1	formula	SSTrt1÷dfTrt1	MS_A/MS_e
Treatment B	b-1	formula	SSTrt2÷dfTrt2	MS_B/MS_e
A×B Interaction	(k-1)(b-1)	formula	SSInt÷dfInt	$MS_{A\times B}/MS_e$
Error	N-kb	formula	SSErr÷dfErr	
Total	N-1	formula	-	

ANOVA in R

Well…maybe a bit of background on how R conducts an ANOVA first….

Completely randomized design

```
> fertilizers <- c("A","B","C")
> replicates <- 7
> assignments <- rep(x=fertilizers,times=replicates)
> assignments
 [1] "A" "B" "C" "A" "B" "C" "A" "B" "C" "A" "B" "C" "A" "B" "C" "A" "B" "C" "A"
[20] "B" "C"
> set.seed(123)
> my_sample <- sample(assignments, replace=F)
> matrix(my_sample, nrow=3, ncol=7)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,] "C"  "C"  "C"  "A"  "A"  "A"  "B"
[2,] "A"  "A"  "B"  "C"  "C"  "C"  "A"
[3,] "B"  "B"  "B"  "C"  "B"  "B"  "A"
```

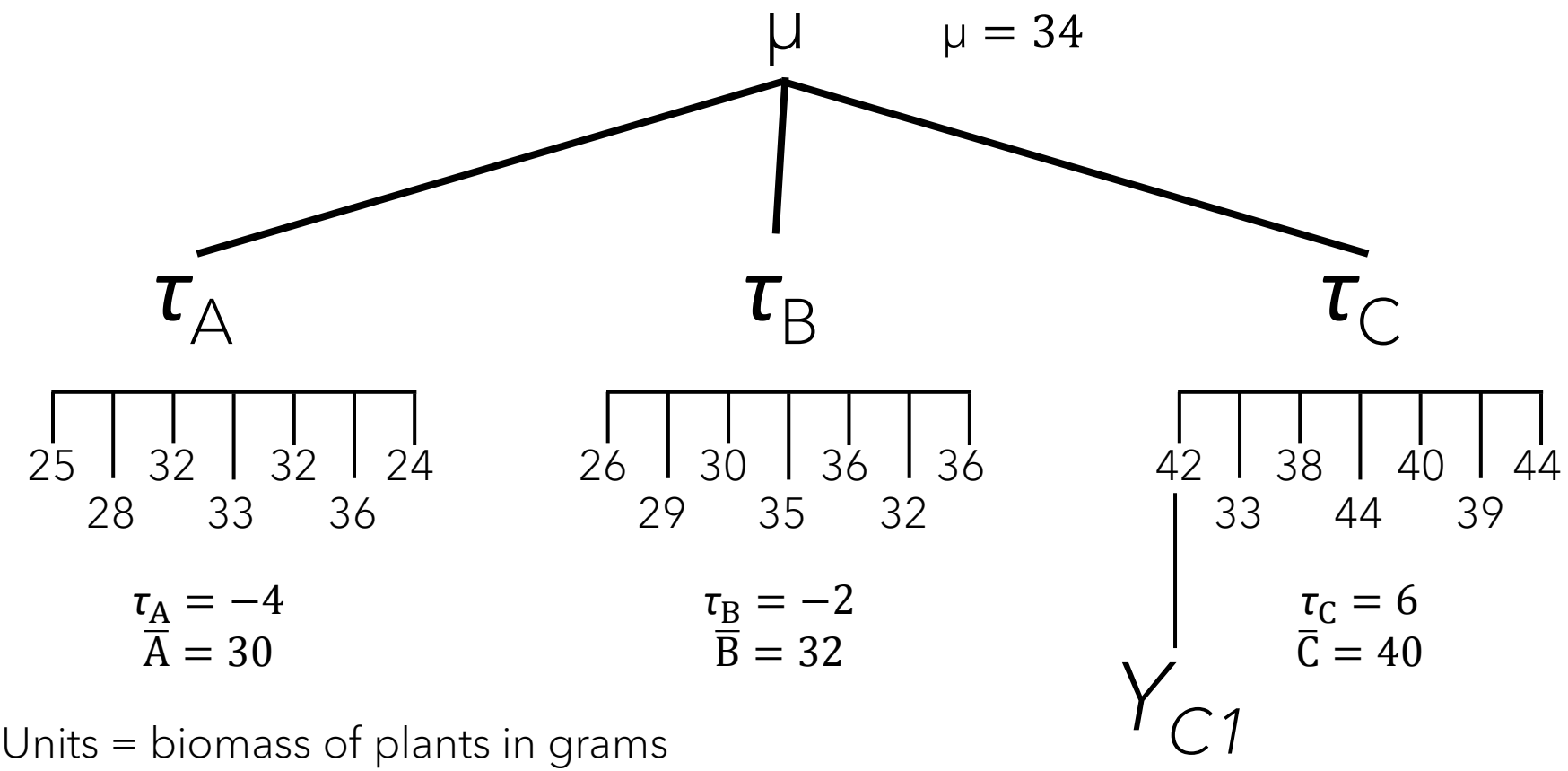


Created by Alexandr Lavreniuk from Noun Project

Linear model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Response Predictors

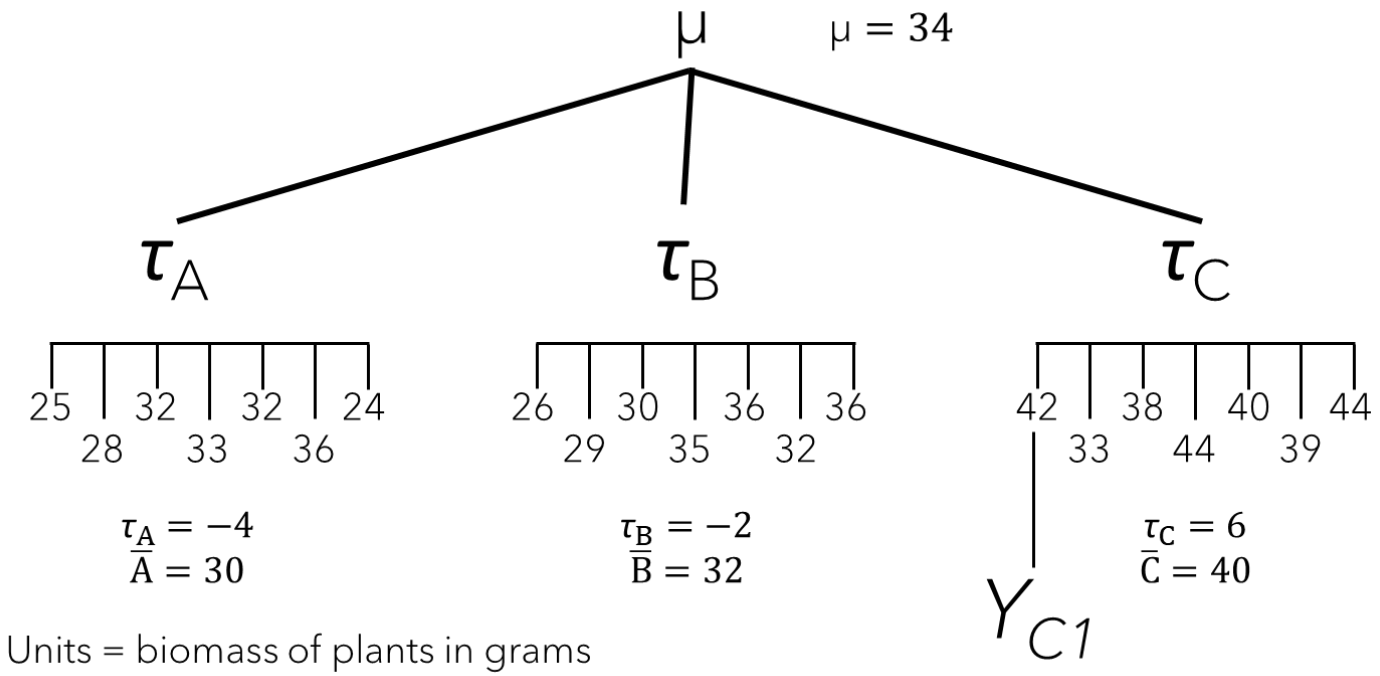


Units = biomass of plants in grams

Activity

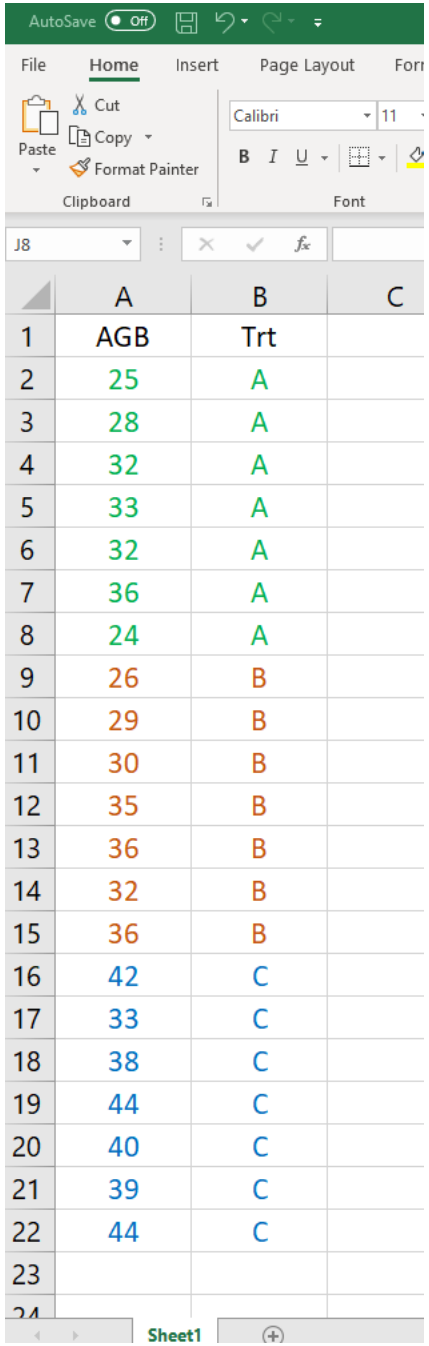
On your scratch paper, organize these data into a spreadsheet ready for analysis in R (the design/layout of a spreadsheet is fine - you don't need to write out the whole thing). How many rows and columns in Excel would you have?

$$\underset{\text{Response}}{Y_{ij}} = \underset{\text{Predictors}}{\mu + \tau_i} + \varepsilon_{ij}$$



Units = biomass of plants in grams

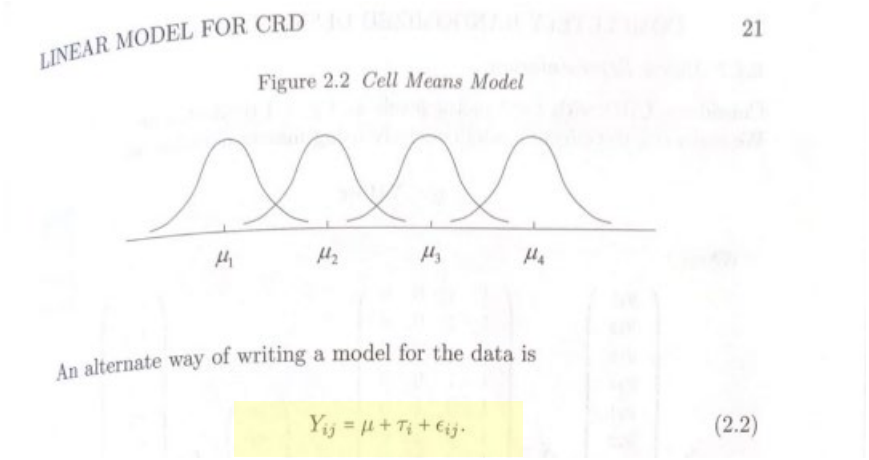
Activity



	A	B	C
1	AGB	Trt	
2	25	A	
3	28	A	
4	32	A	
5	33	A	
6	32	A	
7	36	A	
8	24	A	
9	26	B	
10	29	B	
11	30	B	
12	35	B	
13	36	B	
14	32	B	
15	36	B	
16	42	C	
17	33	C	
18	38	C	
19	44	C	
20	40	C	
21	39	C	
22	44	C	
23			

Linear model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

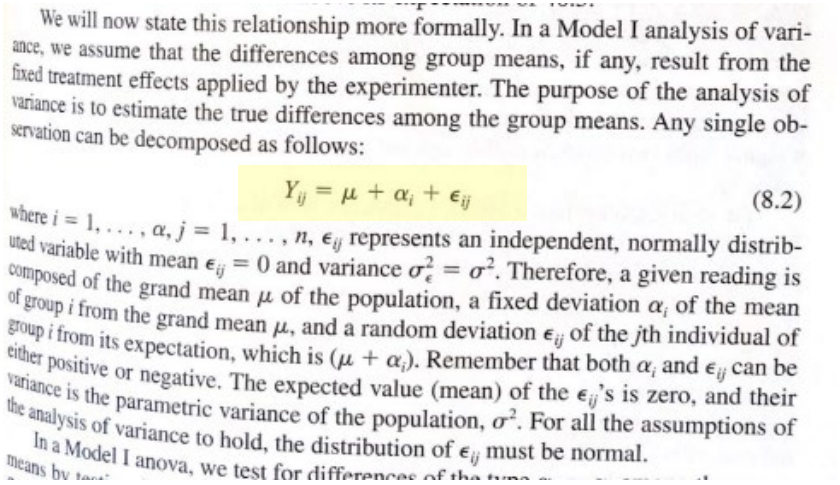


LINEAR MODEL FOR CRD 21

Figure 2.2 *Cell Means Model*

An alternate way of writing a model for the data is

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}. \quad (2.2)$$



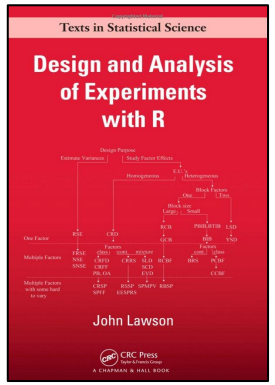
We will now state this relationship more formally. In a Model I analysis of variance, we assume that the differences among group means, if any, result from the fixed treatment effects applied by the experimenter. The purpose of the analysis of variance is to estimate the true differences among the group means. Any single observation can be decomposed as follows:

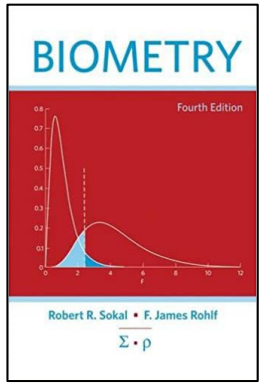
$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad (8.2)$$

where $i = 1, \ldots, \alpha$, $j = 1, \ldots, n$, ϵ_{ij} represents an independent, normally distributed variable with mean $\epsilon_{ij} = 0$ and variance $\sigma_\epsilon^2 = \sigma^2$. Therefore, a given reading is composed of the grand mean μ of the population, a fixed deviation α_i of the mean of group i from the grand mean μ, and a random deviation ϵ_{ij} of the jth individual of group i from its expectation, which is $(\mu + \alpha_i)$. Remember that both α_i and ϵ_{ij} can be either positive or negative. The expected value (mean) of the ϵ_{ij}'s is zero, and their variance is the parametric variance of the population, σ^2. For all the assumptions of the analysis of variance to hold, the distribution of ϵ_{ij} must be normal.

In a Model I anova, we test for differences of the type

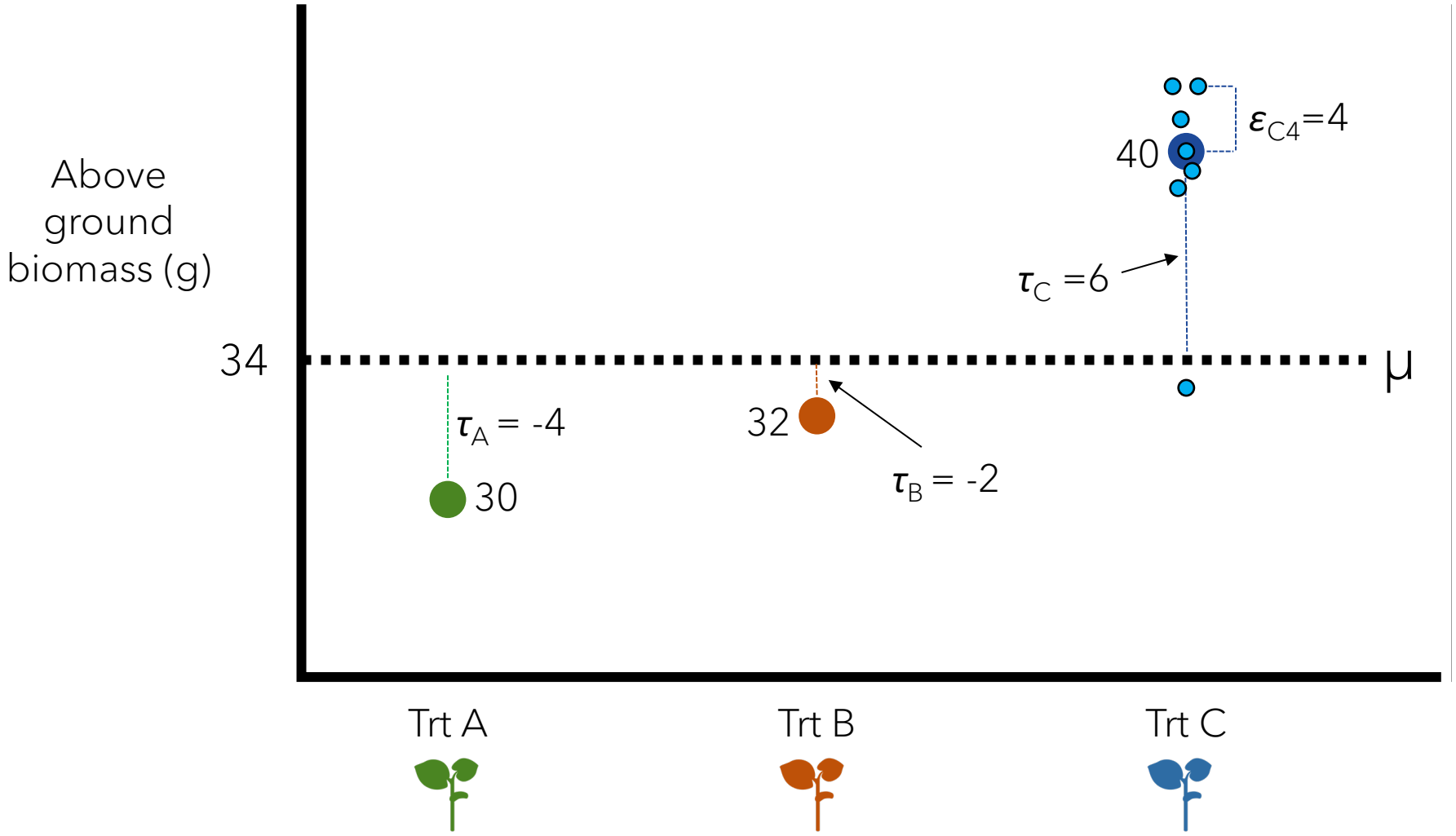
means by





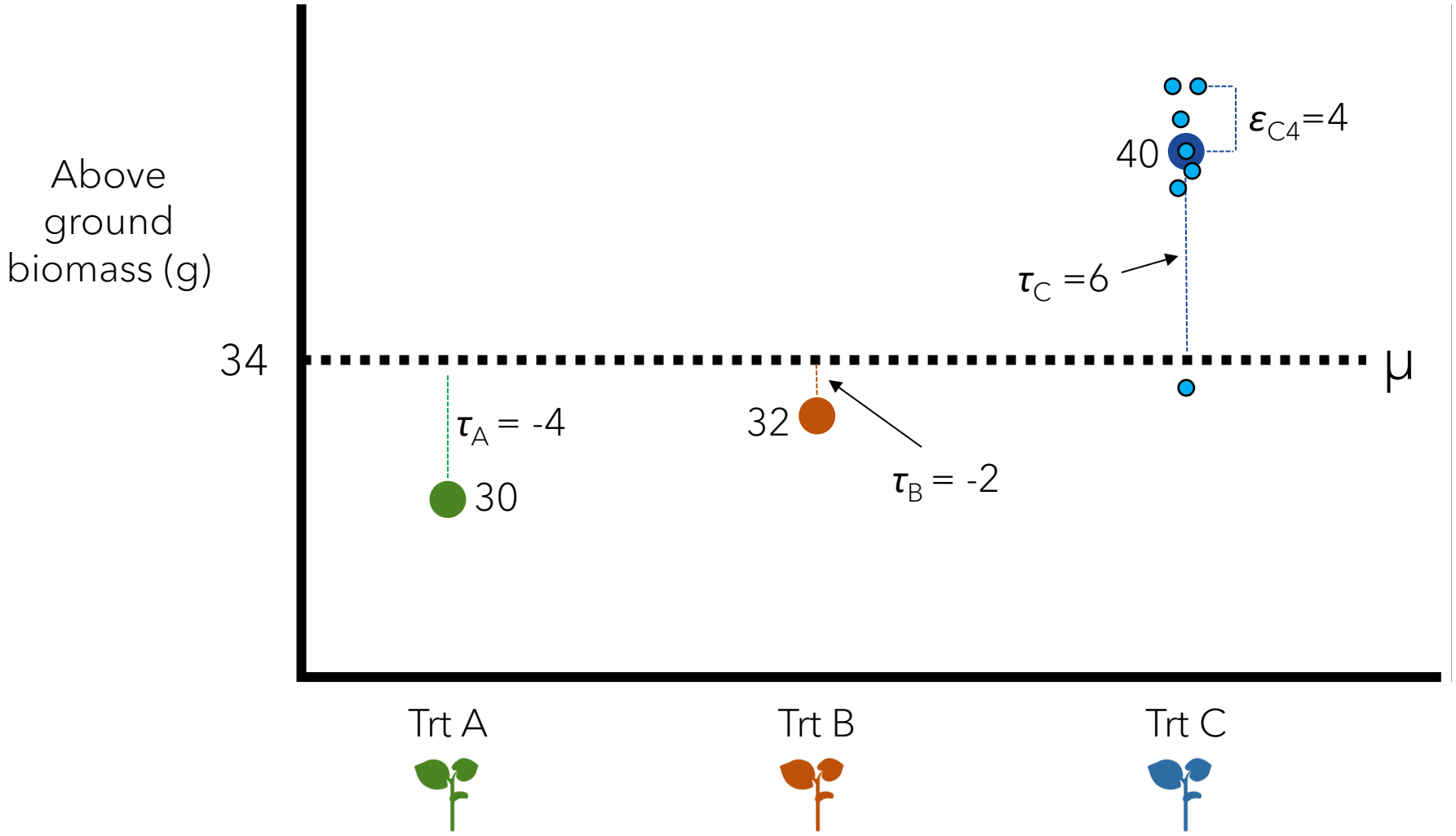
Linear model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$



Linear model

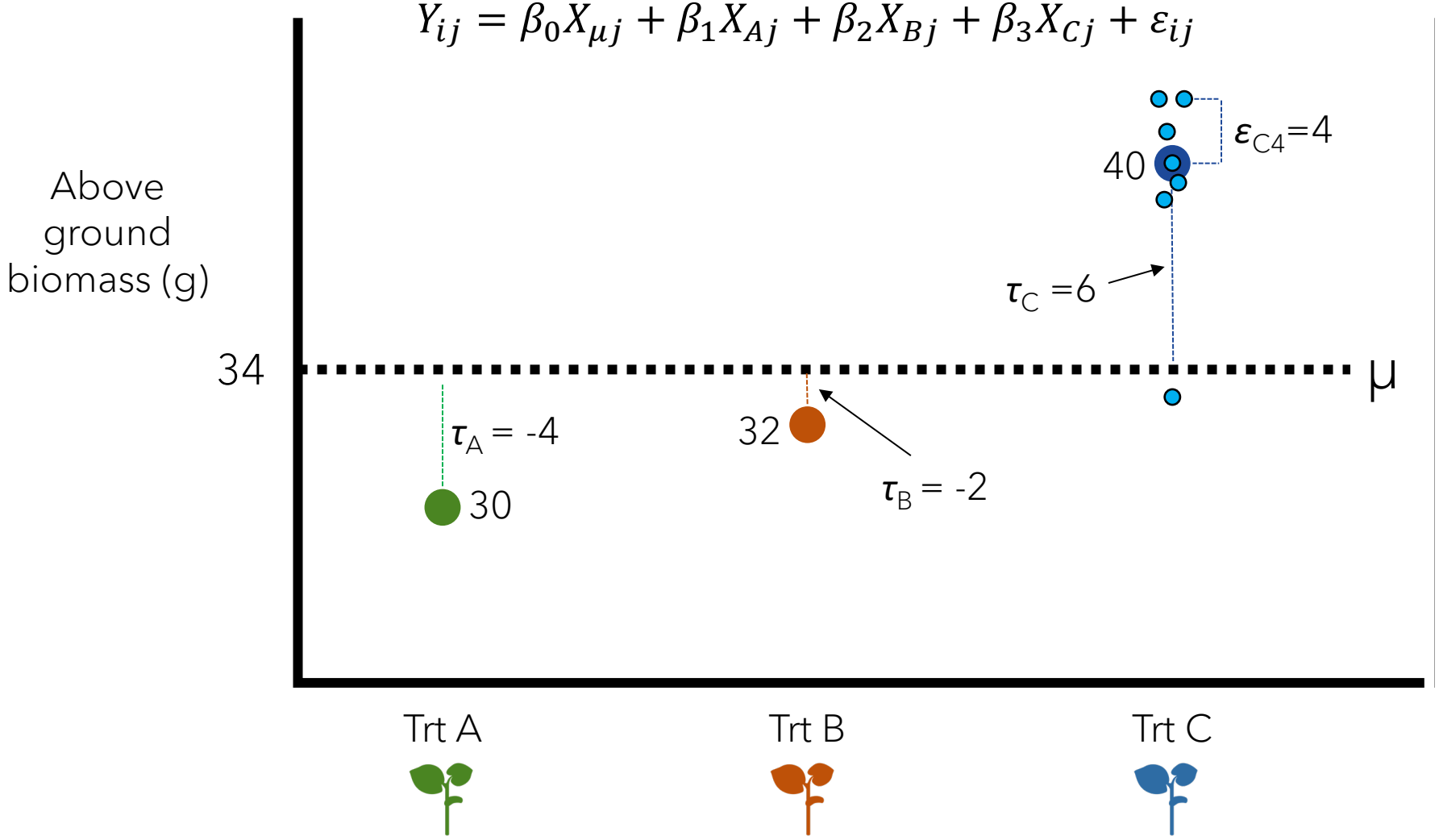
$$Y_{cj} = 34 + 6_c + \varepsilon_{ij}$$



Linear model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

$$Y_{ij} = \beta_0 X_{\mu j} + \beta_1 X_{Aj} + \beta_2 X_{Bj} + \beta_3 X_{Cj} + \varepsilon_{ij}$$

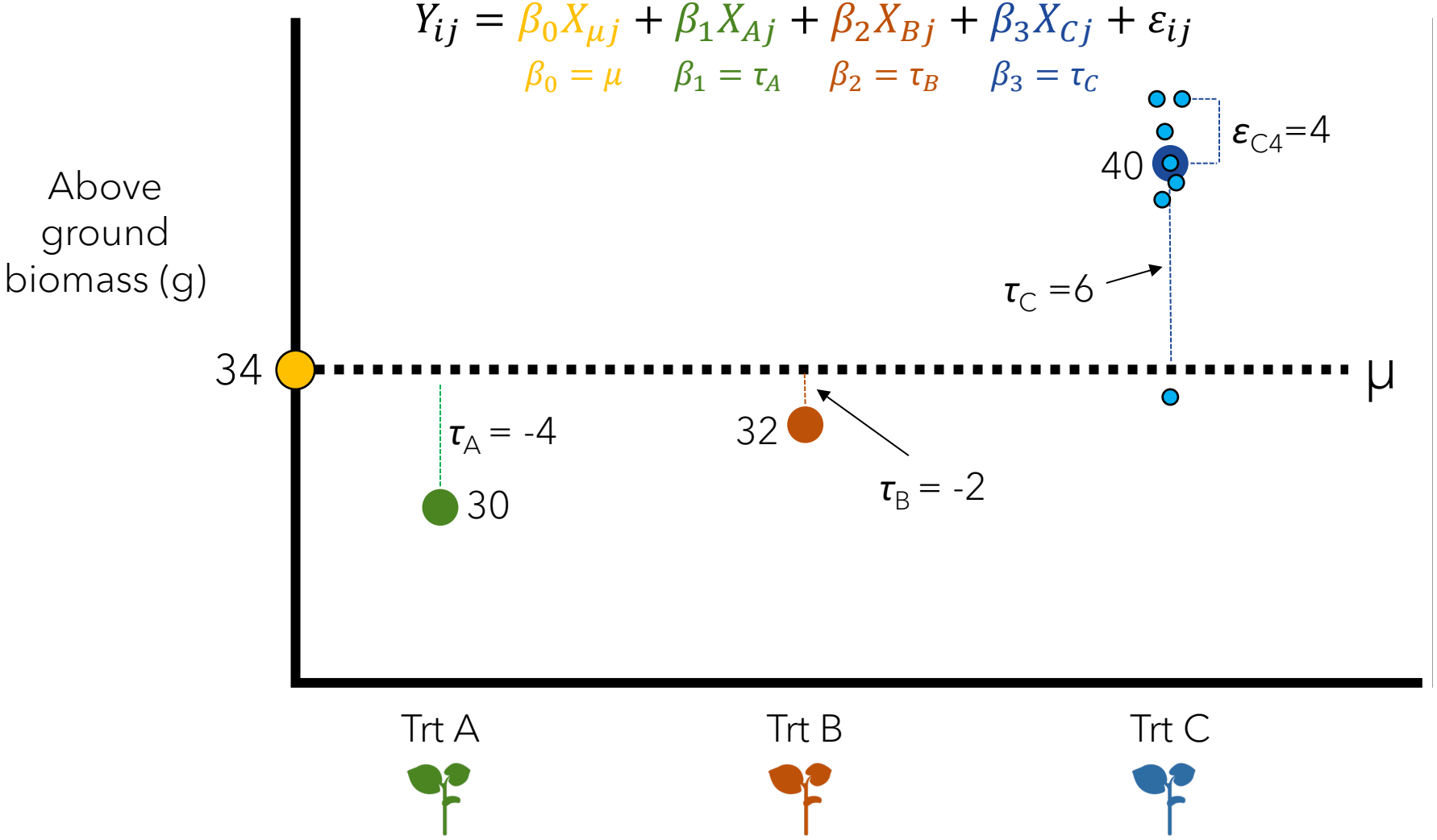


Linear model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

$$Y_{ij} = \beta_0 X_{\mu j} + \beta_1 X_{Aj} + \beta_2 X_{Bj} + \beta_3 X_{Cj} + \varepsilon_{ij}$$

$\beta_0 = \mu$ $\beta_1 = \tau_A$ $\beta_2 = \tau_B$ $\beta_3 = \tau_C$



Linear models in R

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

$$Y_{ij} = \beta_0 X_{\mu j} + \beta_1 X_{Aj} + \beta_2 X_{Bj} + \beta_3 X_{Cj} + \varepsilon_{ij}$$

$$\beta_0 = \mu \qquad \beta_1 = \tau_A \qquad \beta_2 = \tau_B \qquad \beta_3 = \tau_C$$

$$Y_{ij} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix}, X_{ij} = \begin{array}{c} \begin{array}{cccc} X_\mu & X_A & & \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \end{array}, \beta = \begin{pmatrix} \mu \\ \tau_A \\ \tau_B \\ \tau_C \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \\ \varepsilon_{34} \end{pmatrix}$$

Linear models in R

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

$$Y_{ij} = \beta_1 X_{Aj} + \beta_2 X_{Bj} + \beta_3 X_{Cj} + \varepsilon_{ij}$$

$$Y_{ij} = \beta_0 + \beta_1 X_{Bj} + \beta_2 X_{Cj} + \varepsilon_{ij}$$

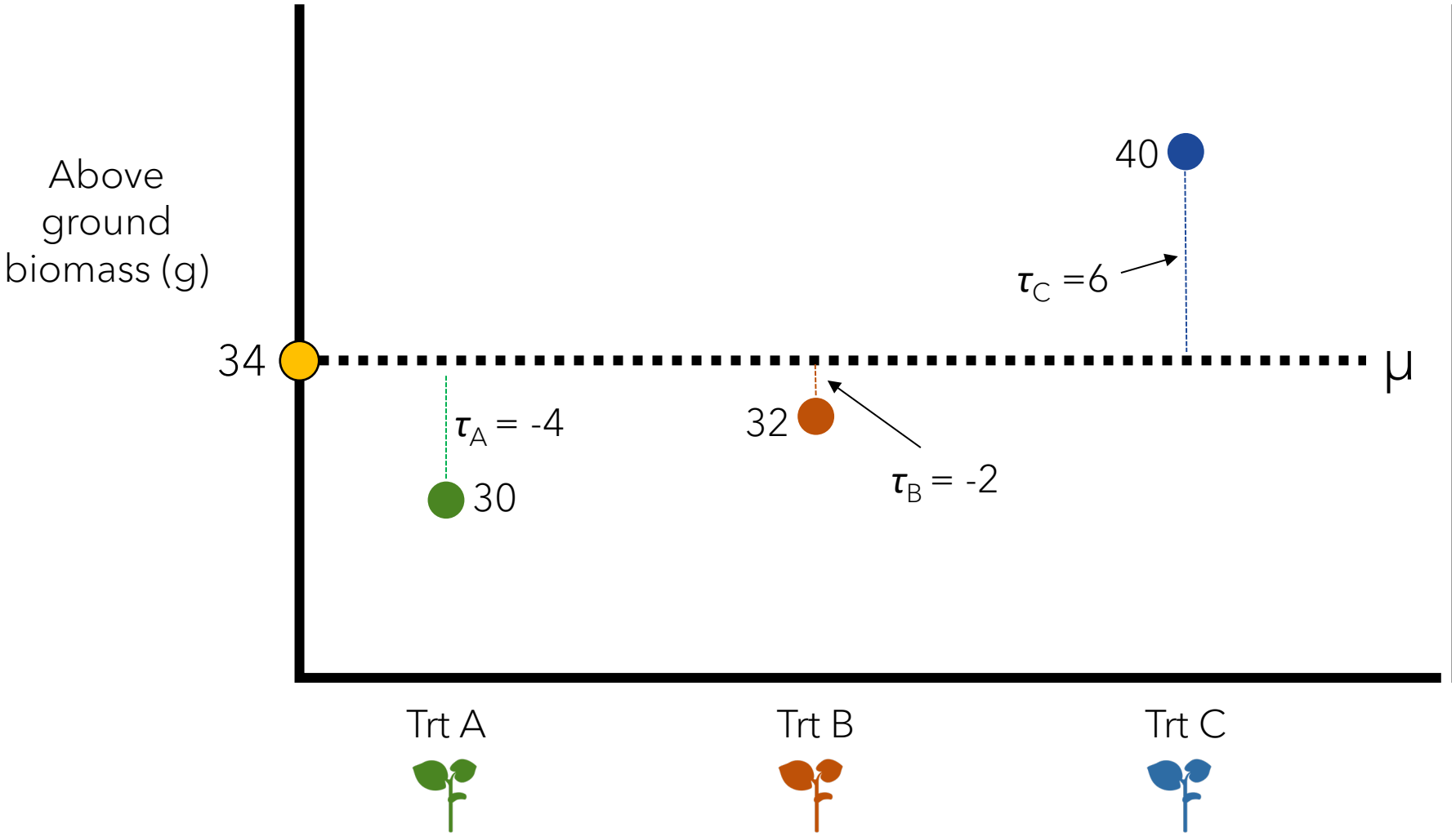
$$\boldsymbol{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \beta = \begin{pmatrix} \overbrace{\tau_B - \tau_A}^{\substack{\beta_1 \\ \mathbf{2} = (-2) - (-4)}} \\ \underbrace{\tau_C - \tau_A}_{\substack{\mathbf{10} = (6) - (-4) \\ \beta_2}} \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \\ \varepsilon_{34} \end{pmatrix}$$

Linear models in R

	A	B	C	D	E	F	G
1	AGB	Trt		Y	TrtA	TrtB	TrtC
2	25	A		25	1	0	0
3	28	A		28	1	0	0
4	32	A		32	1	0	0
5	33	A		33	1	0	0
6	32	A		32	1	0	0
7	36	A		36	1	0	0
8	24	A		24	1	0	0
9	26	B		26	1	1	0
10	29	B		29	1	1	0
11	30	B		30	1	1	0
12	35	B		35	1	1	0
13	36	B		36	1	1	0
14	32	B		32	1	1	0
15	36	B		36	1	1	0
16	42	C		42	1	0	1
17	33	C		33	1	0	1
18	38	C		38	1	0	1
19	44	C		44	1	0	1
20	40	C		40	1	0	1
21	39	C		39	1	0	1
22	44	C		44	1	0	1
23							

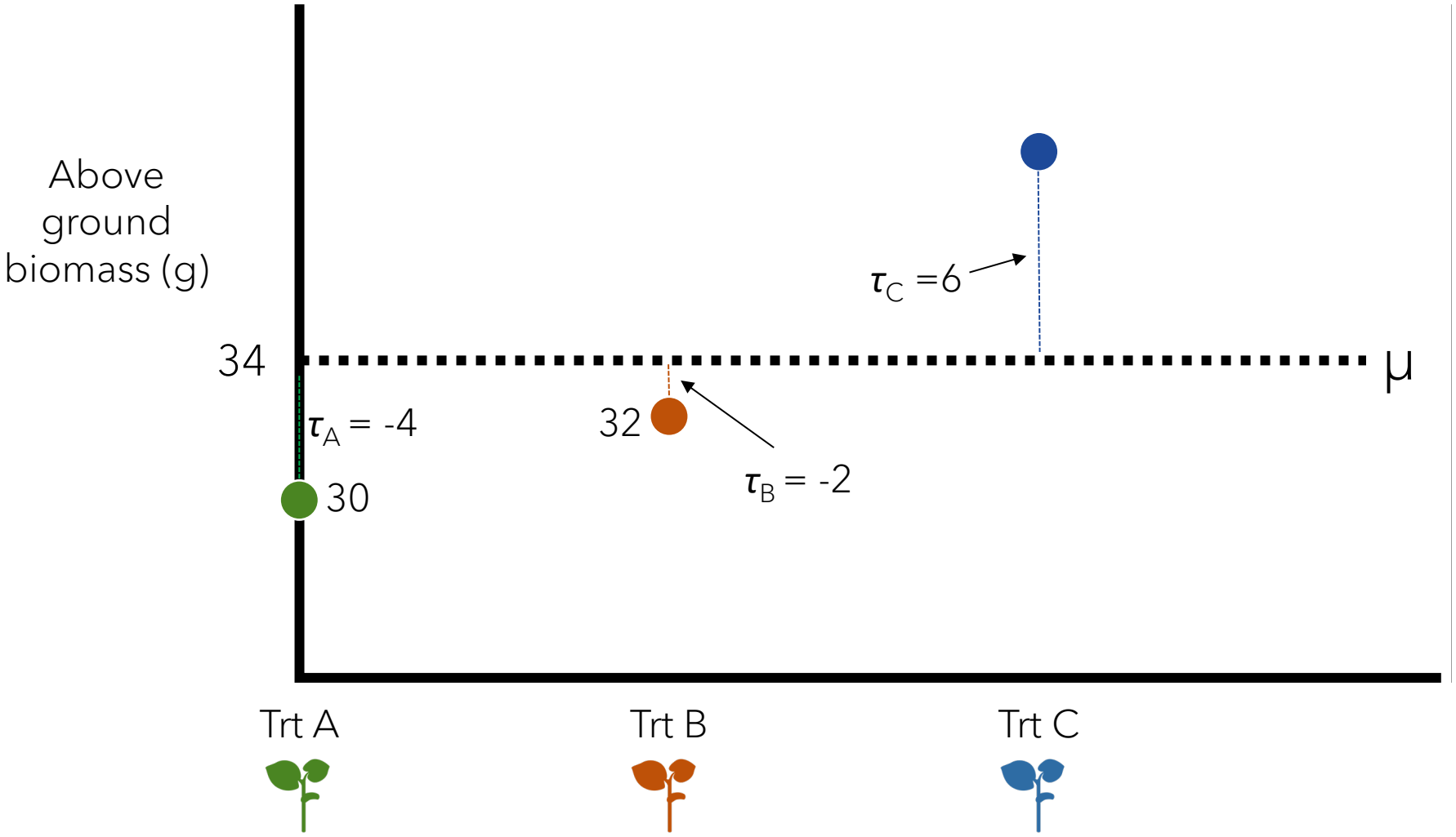
Linear models in R

$$Y_{ij} = \beta_0 X_{\mu j} + \beta_1 X_{Aj} + \beta_2 X_{Bj} + \beta_3 X_{Cj} + \varepsilon_{ij}$$



Linear models in R

$$Y_{ij} = \beta_0 + \beta_1 X_{Bj} + \beta_2 X_{Cj} + \varepsilon_{ij}$$

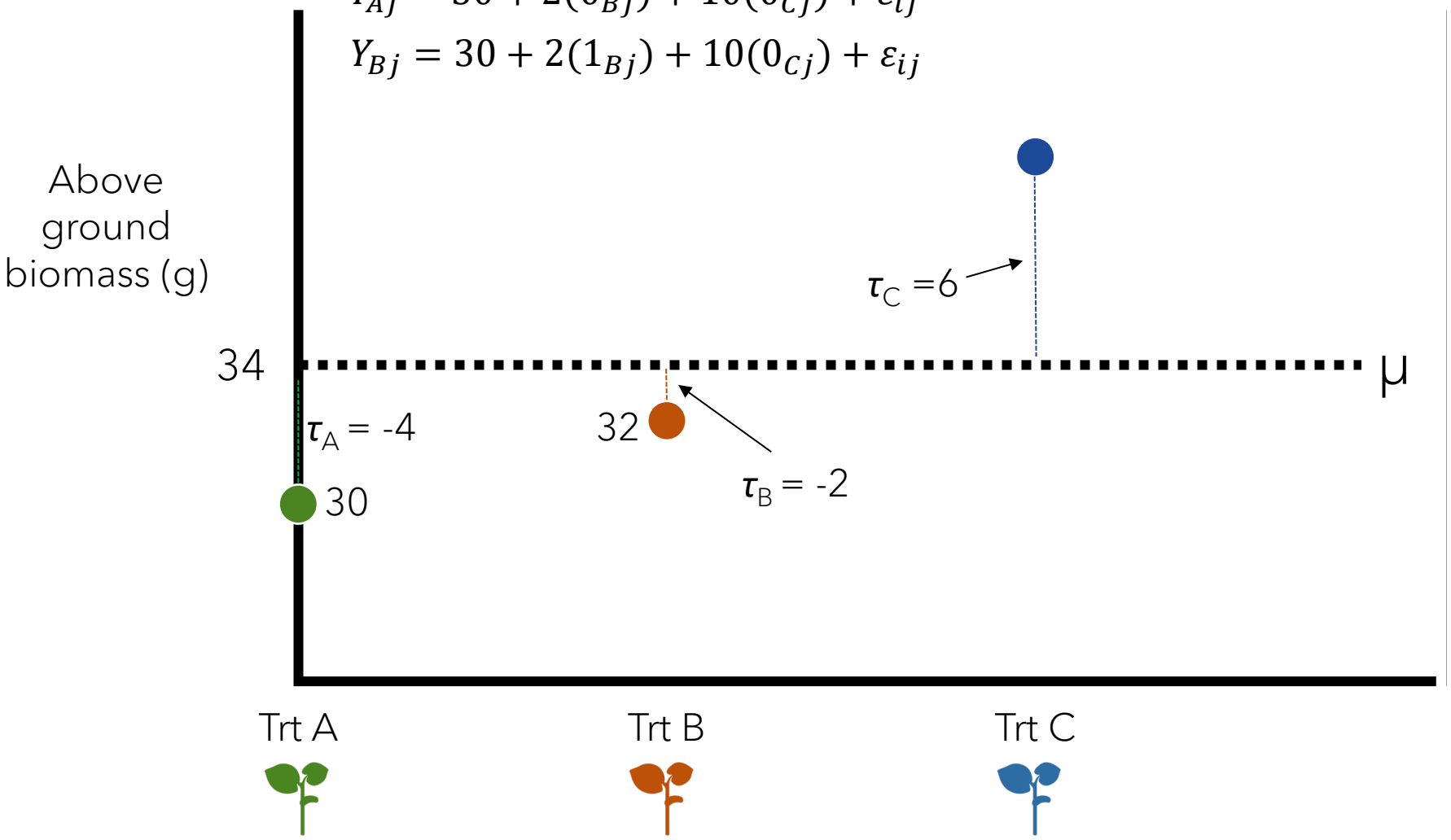


Linear models in R

$$Y_{ij} = \beta_0 + \beta_2 X_{Bj} + \beta_3 X_{Cj} + \varepsilon_{ij}$$

$$Y_{Aj} = 30 + 2(0_{Bj}) + 10(0_{Cj}) + \varepsilon_{ij}$$

$$Y_{Bj} = 30 + 2(1_{Bj}) + 10(0_{Cj}) + \varepsilon_{ij}$$



Linear models in R

```
> SB_df
   AGB Trt
1   25   A
2   28   A
3   32   A
4   33   A
5   32   A
6   36   A
7   24   A
8   26   B
9   29   B
10  30   B
11  35   B
12  36   B
13  32   B
14  36   B
15  42   C
16  33   C
17  38   C
18  44   C
19  40   C
20  39   C
21  44   C
```

fit1 <- lm(AGB~Trt, data=SB_df)

```
> anova(fit1)
Analysis of Variance Table

Response: AGB
          Df Sum Sq Mean Sq F value    Pr(>F)
Trt        2    392 196.000  11.839 0.0005228 ***
Residuals 18    298  16.556
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
>
> # linear model (regression)
> summary(fit1)

Call:
lm(formula = AGB ~ Trt, data = SB_df)

Residuals:
   Min     1Q Median     3Q    Max
    -7     -2      0      3      6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   30.000      1.538  19.507 1.48e-13 ***
TrtB           2.000      2.175   0.920 0.369949
TrtC          10.000      2.175   4.598 0.000223 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.069 on 18 degrees of freedom
Multiple R-squared:  0.5681,    Adjusted R-squared:  0.5201
F-statistic: 11.84 on 2 and 18 DF,  p-value: 0.0005228
```