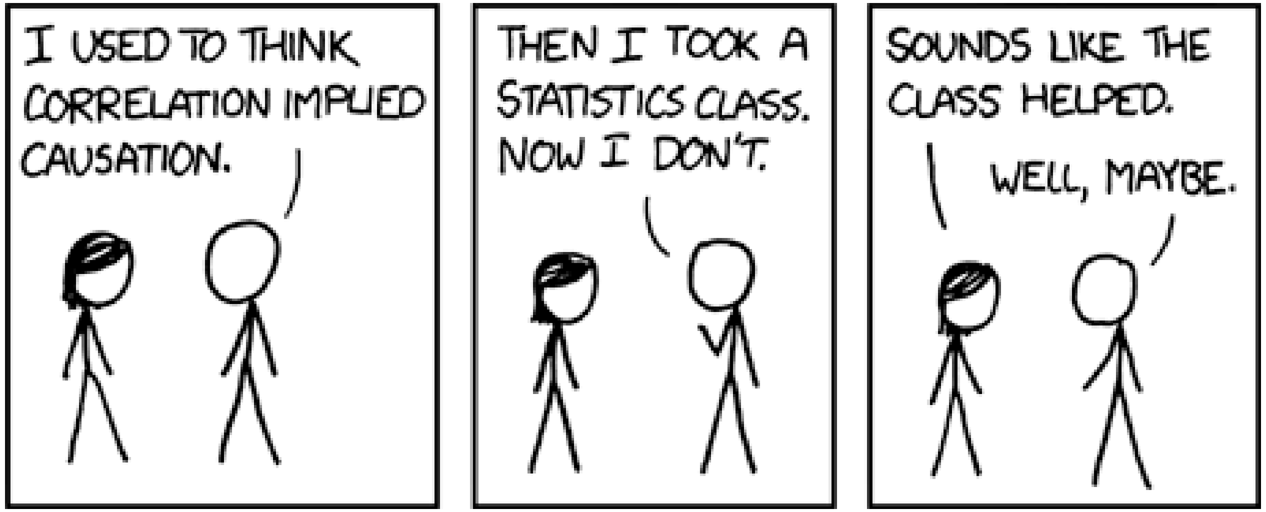
Introduction to Experimental Design

ENTMLGY 6707 Entomological Techniques and Data Analysis


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Questions to think about while we wait to get started:

- What are some potential "target journals" where your research could be published one day?
- Pick one of those journals. What is their policy on "open science"? Do they require code and data sharing? Feel free to Google away while we wait!

Learning objectives

1. Define key steps and terms in experimental design
2. Compare and contrast methods of treatment allocation
3. Practice writing and using a linear model
4. Discuss potential actions following a failed experiment
5. Evaluate advantages and disadvantages of observational studies

Some vocabulary

Null hypothesis (H_0): (1) measurements – most commonly we are interested in means/averages - do NOT differ significantly among or between treatments or populations; any observed differences arise due to random chance (e.g., sampling, experimental noise). (2) measurements do not differ significantly from a set value (e.g., 0).

Alternative hypothesis (H_a): (1) measurements (means) differ significantly among or between treatments or populations. (2) measurements are significantly different from a set value (e.g., 0).

Experimental Unit: the smallest entity to which a treatment is applied

Unit of observation: the unit at which data are collected (you will sometimes see "sampling unit" used equivalently)

Unit of analysis: the unit at which data are analyzed (frequently – but not always – the same as the unit of observation)

Apply it

Null hypothesis (H_0): (1) measurements – most commonly we are interested in means/averages - do NOT differ significantly among or between treatments or populations; any observed differences arise due to random chance (e.g., sampling, experimental noise). (2) measurements do not differ significantly from a set value (e.g., 0).

Alternative hypothesis (H_a): (1) measurements (means) differ significantly among or between treatments or populations. (2) measurements are significantly different from a set value (e.g., 0).

Experimental Unit: the smallest entity to which a treatment is applied

Unit of observation: the unit at which data are collected (you will sometimes see "sampling unit" used equivalently)

Unit of analysis: the unit at which data are analyzed (frequently – but not always – the same as the unit of observation)

Activity: Think of one experiment or study from your thesis (completed or planned). What are the hypotheses, experimental unit, and unit of observation?

Please write this up (bullet points are fine) and submit to Check-in 3 on Canvas

Experimental design: key principles

For each term:

- Explain its importance in designing robust experiments.
- Explain how each would be achieved “in the real world.”

Randomization
Replication
Reduction of noise

Experimental design: key principles

Randomization: the random allocation of treatments to experimental units.

- Allocation should actually be based on something random. That is, haphazard ≠ random.
- Protect against biases
- Helps one meet assumptions of independence
  → enables statistical analysis
- Note that treatments may include things that are categorical (e.g., presence of thinning or an insecticide) OR continuous (e.g., target basal area or application rate).

Replication: the repetition of a treatment within an experiment.

- Reduce the effect of uncontrolled variation
- Increase precision
- Quantify uncertainty

Reduction of noise: by controlling as much as possible the conditions in the experiment.

Experimental design (Hurlbert 1984)

Specify:

1. Nature of the experimental units
2. Numbers and kinds of treatments to be imposed
3. Properties or responses that will be measured
4. Manner in which treatments are assigned to experimental units
5. Number of experimental units receiving each treatment (replicates)
6. Physical (spatial) arrangement of units
7. Temporal sequence of treatment application

Types of experiments

Mensurative: taking measurements at one or more points in time or space; no imposition of experimental treatments. The “treatments” or “predictors” are implemented by taking advantage of spatiotemporal variation.

- Sometimes (usually?) called observational studies or observational experiments.

Manipulative: certain experimental units receive different treatments.

Experimental design: the variables

Response variable: aka dependent variable. A random variable whose value depends on that of another. Often represented using *y or* **Y**. Measured at your unit of observation.

Predictor variable: aka independent or explanatory variable. You select “treatments” to be measured at experimental units (mensurative) or assign treatments of some primary factor to experimental units (manipulative). Determine if and what controls are necessary. Often represented using x *or* ***X***.

“Nuisance” variable: random variable that is of no particular interest but that likely affects your response variable (potentially for reasons unknown).

Example R code for making a graph and fitting a linear model:

```
plot(y~x, data= df) # I would read this “plotting y as a function of x”
plot(y ~ x, data= df) # I would read this “plotting y as a function of x”
fit1 <- lm(y~x, data= df) # I would read this “modeling y as a function of x”

plot(response_variable1 ~ treatments, data=df_trt)
fit1 <- lm(response_variable1 ~ treatments, data=df_trt)

plot(species_richness ~ latitude, data=df_trt)
fit1 <- lm(species_richness ~ latitude, data=df_trt)
```

Controls

In a broad sense, control means to reduce the amount of noise or variation (e.g., get instruments serviced, avoid or account for natural gradients in experiments).

Control groups (vs. treatment groups)
Untreated: no imposition of an experimental variable
Treated: procedural treatment
Positive control: a control in which you expect a response
Negative control: a control in which you do not expect a response

The goal is to isolate - as much as possible – the effect of the treatment by accounting for confounding variables (e.g., procedural or temporal changes in your response).

Exercise: You are interested in growth rates of a newly discovered plant pathogenic fungi on different media. You plan to inoculate each medium by transferring the spores from plants using a cotton swab. Can you think of positive and negative controls you might want to include?

Controls

When designing your control(s), think hard about the biology AND experimental protocols (e.g., does the design require several manipulations that might have unwanted effects?).

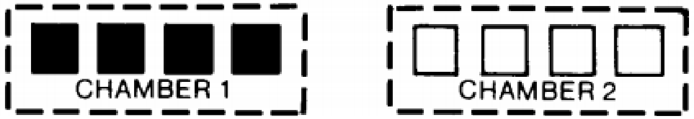
Hannah Locke @HLockeSTEM · 14h

I DO THIS TO MYSELF

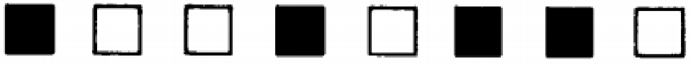
ME DESIGNING INCREDIBLY AMBITIOUS EXPERIMENTS
ME EXECUTING INCREDIBLY AMBITIOUS EXPERIMENTS
imgflip.com
17
160
1.3K

Allocating treatments

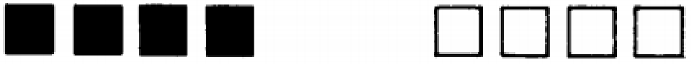
A.



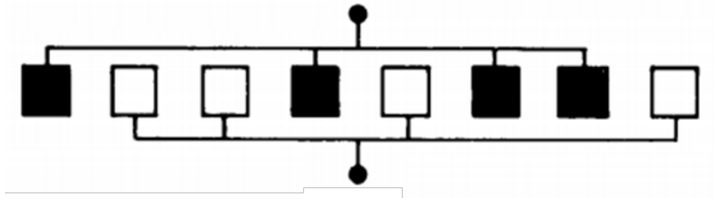
B.



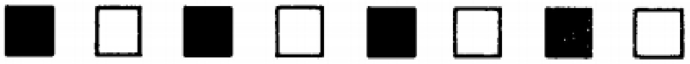
C.



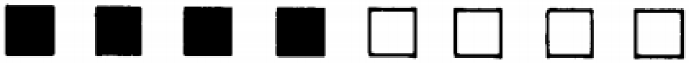
D.



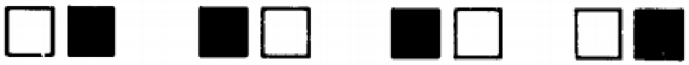
E.



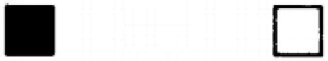
F.



G.



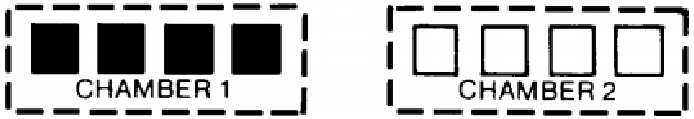
H.



1. Isolative segregation
2. No replication
3. Randomized, but with inter-dependent replicates
4. Randomized block
5. Clumped segregation
6. Completely randomized
7. Systematic
8. Simple segregation

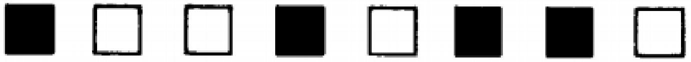
Allocating treatments

A.



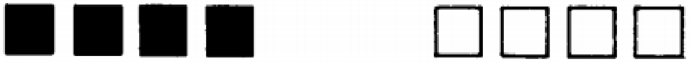
Isolative segregation

B.



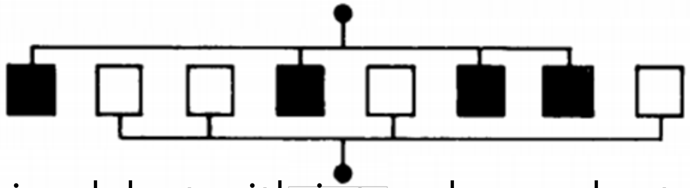
Completely randomized

C.



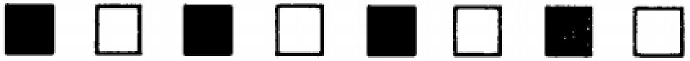
Clumped segregation

D.



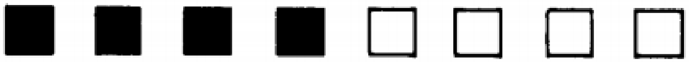
Randomized, but with inter-dependent replicates

E.



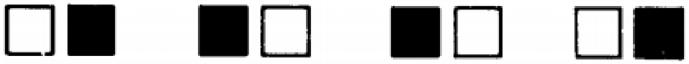
Systematic

F.



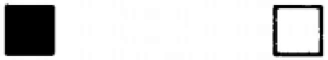
Simple segregation

G.



Randomized block

H.



No replication

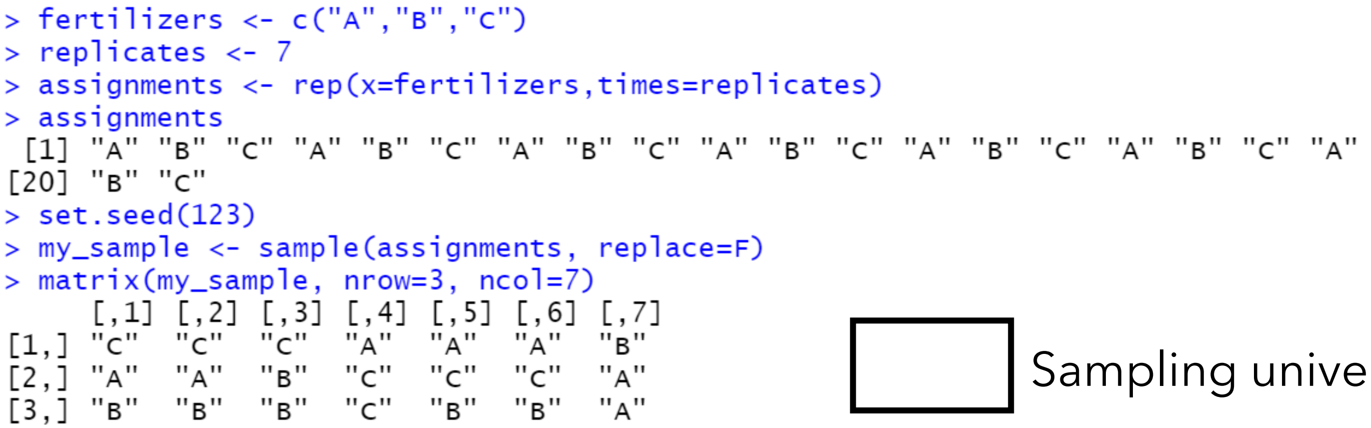
Part 2: Identify which ones you would feel comfortable using. Be prepared to explain your reasoning.

Simple experiment

Measure growth of plants (biomass of individuals, g) as a function of three different fertilizers (A = none/control, B = Ammonium sulfate, and C = Urea). The fertilizer will be applied to a 20×20cm^2 plot around the base of each plant in spring (assume no interference between fertilizers on adjacent plants). In autumn, the above ground biomass (AGB) of each plant will be cut, dried, and weighed.

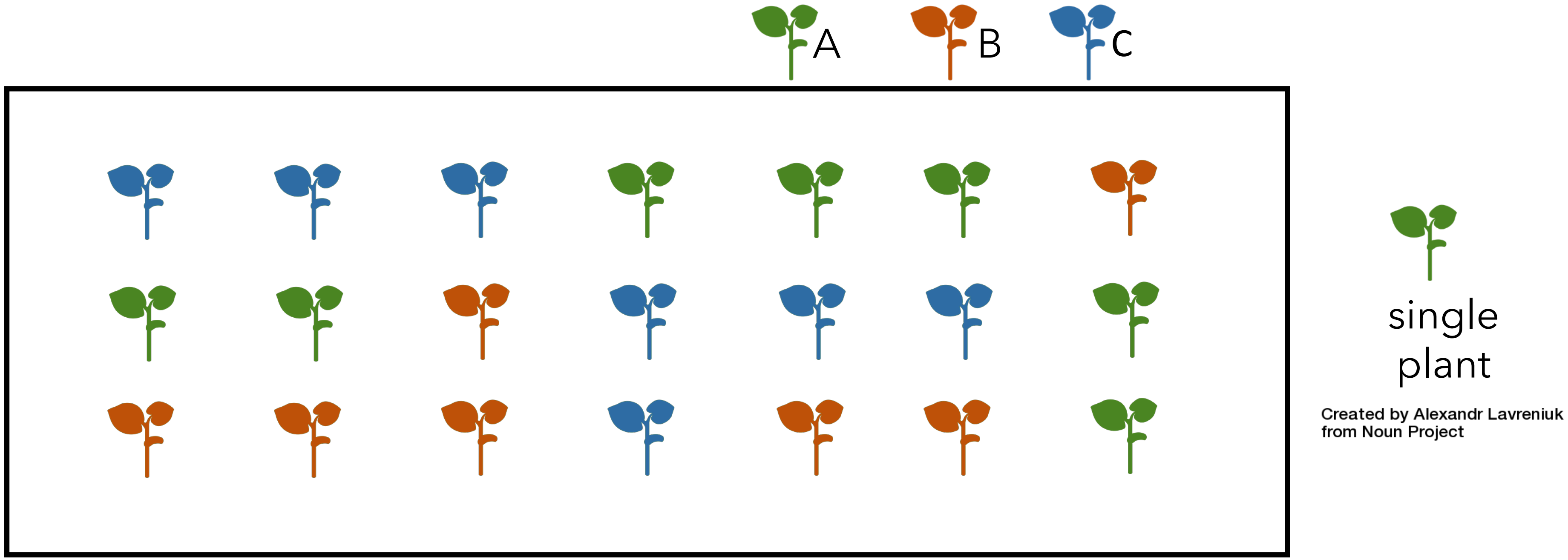
Identify:
Experimental Unit
Unit of observation
Null hypothesis
Alternative hypothesis (or at least one potential one)

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" 
```

□ Sampling universe (e.g., farm)



Created by Alexandr Lavreniuk from Noun Project

Quick refresher on notation

$$Y = (4,8,1,1,30,5)$$

$$Y_1 = 4$$
$$Y_5 = 30$$

Note that you will use (or have already used) similar notation in R, for example, when referring to elements in a vector:

```
> vector_subset <- vector_full[1:5] # get first five items
```

Source for adorable broccoli: StickerApp

Linear model

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

Response Predictors

Y_{ij} is the value for the *j*th observation in the *i*th level of the treatment.
μ is the overall mean (average) across all observations
τ_i is the effect, or the difference between the *i*th level of the treatment τ and the overall average (μ)
ε_{ij} is the error, or the remaining difference between the *j*th observation and the mean of the *i*th treatment

Linear model

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

Response Predictors

Y_{ij} is the value for the *j*th observation in the *i*th level of the treatment.
μ is the overall mean (average) across all observations
τ_i is the effect, or the difference between the *i*th level of the treatment τ and the overall average (μ)
ε_{ij} is the error, or the remaining difference between the *j*th observation and the mean of the *i*th treatment

For our plant growth experiment, let's say the overall mean AGB was 34g and plants in treatments A, B, and C had average AGBs of 30g, 32g, and 40g, respectively. The biomass of the first plant in treatment C was 42g (Y_{C1} = 42).

Exercise: With this information, find ε_{C1}

$\mu = 34$
$\tau_A = 30-34=-4$
$\tau_B = 32-34=-2$
$\tau_C = 40-34=6$

$Y_{C1} = \mu + \tau_C + \varepsilon_{C1}$
$42 = 34 + 6 + \varepsilon_{C1}$
$\varepsilon_{C1} = 2$
$\varepsilon_{3,1} = 2$
ε_{ij} = observed-expected = $Y_{ij}-(\mu+\tau_i)$

Linear model

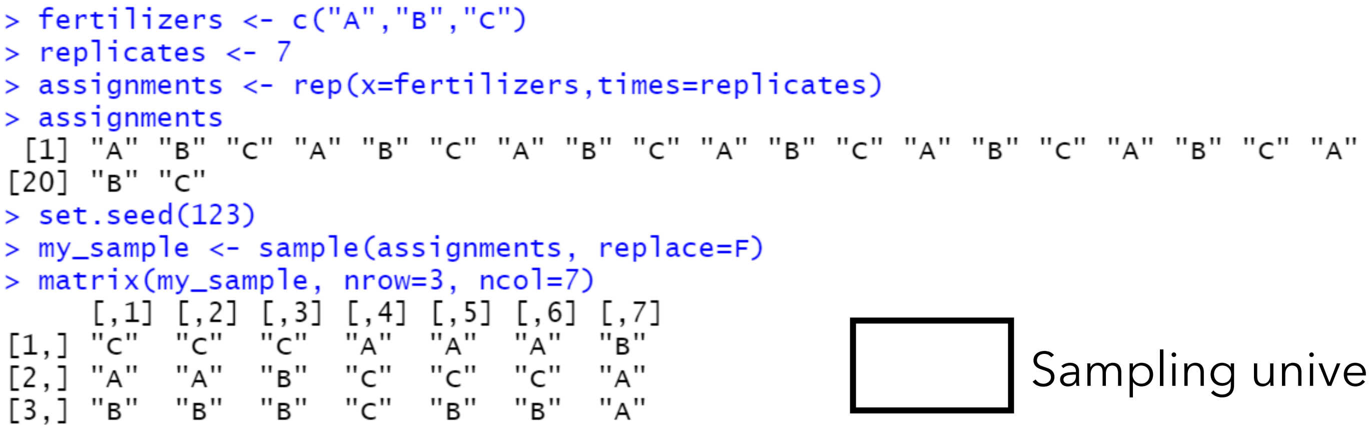
$$\underbrace{Y_{ij}}_{\text{Response}} = \underbrace{\mu + \tau_i}_{\text{Predictors}} + \varepsilon_{ij}$$

Note that we can also write a linear model as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

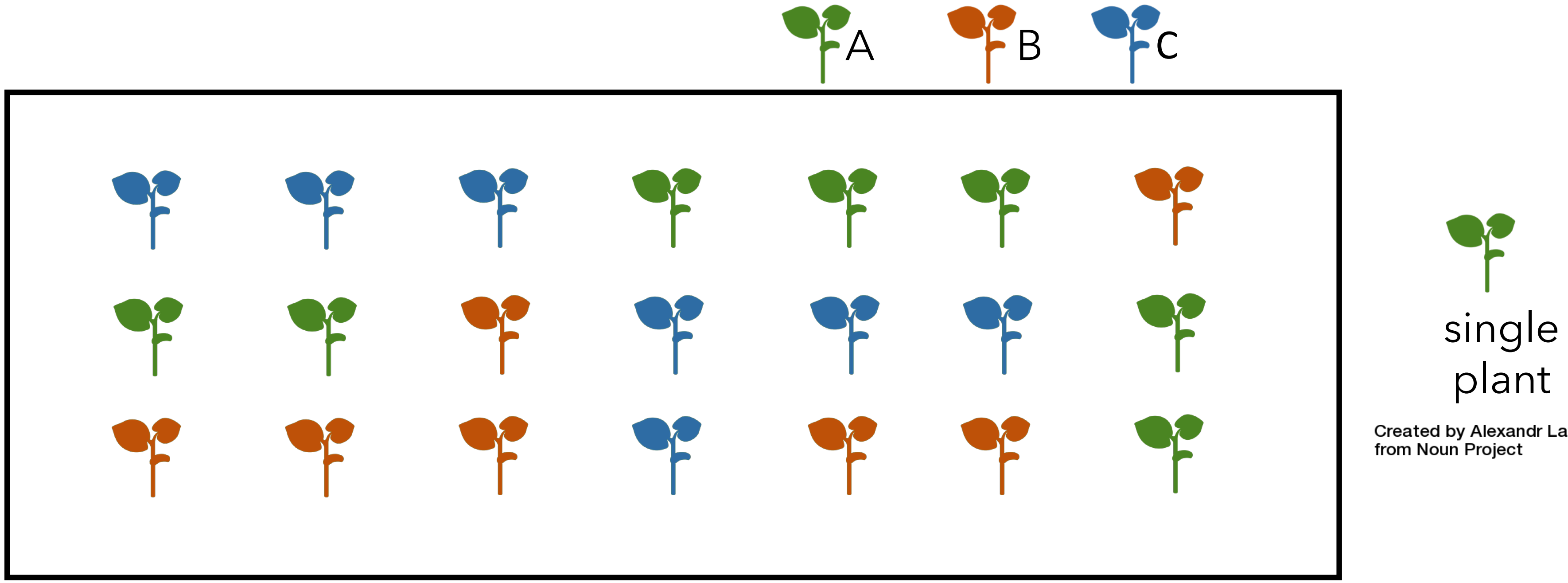
And we will be spending time – a minimal amount – with both notations this semester.

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" 
```

□ Sampling universe (e.g., farm)

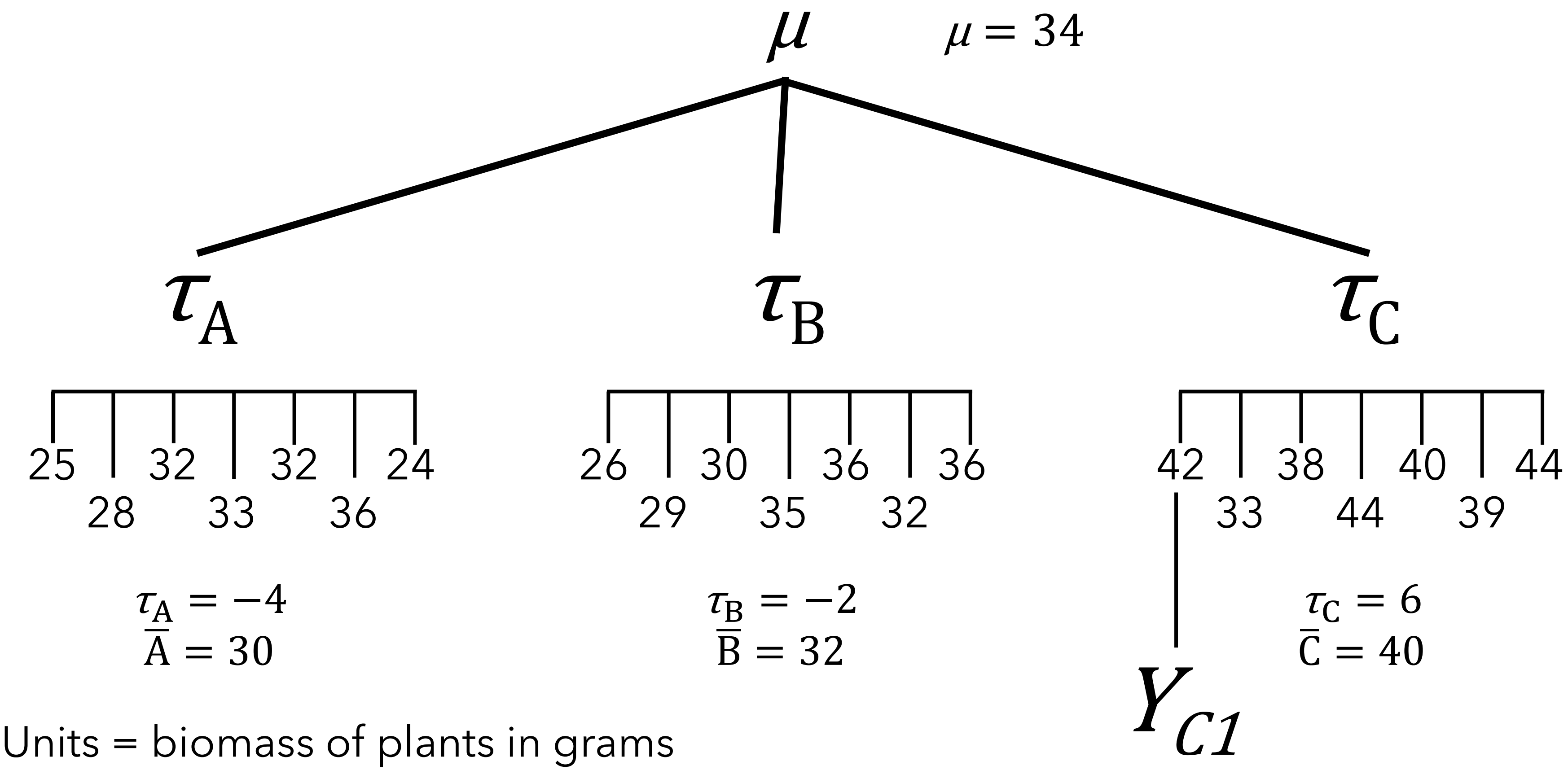


Created by Alexandr Lavreniuk from Noun Project

Linear model

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

Response Predictors



μ

$\mu = 34$

τ_{A}

25 28 32 33 32 36 24

$\tau_{\text{A}} = -4$
$\bar{\text{A}} = 30$

τ_{B}

26 29 30 35 36 32 36

$\tau_{\text{B}} = -2$
$\bar{\text{B}} = 32$

τ_{C}

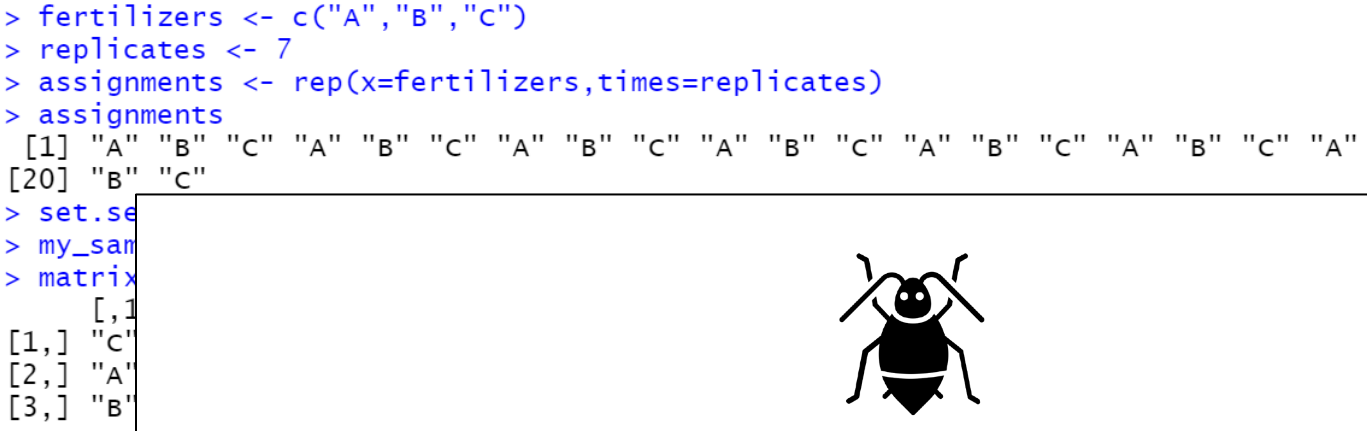
42 33 38 44 40 39 44

Y_{C1}

$\tau_{\text{C}} = 6$
$\bar{\text{C}} = 40$

Units = biomass of plants in grams

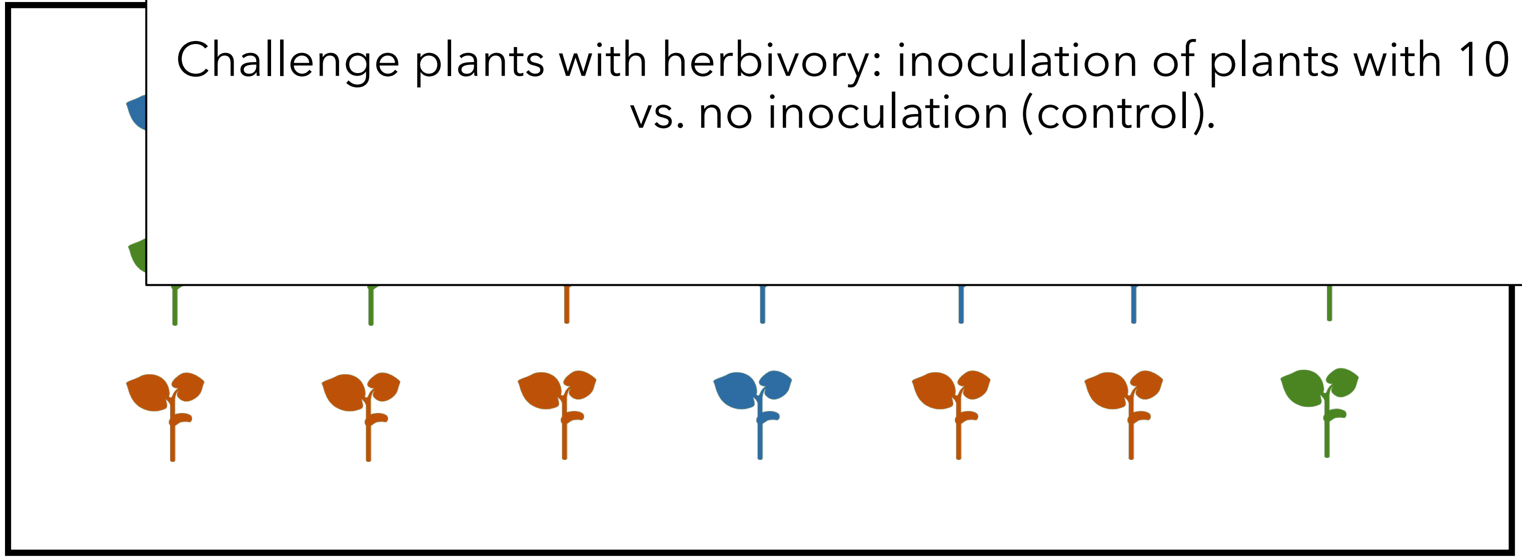
Completely randomized design



```
> 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.se
> my_sam
> matrix
     [,1
[1,] "C"
[2,] "A"
[3,] "B"
```

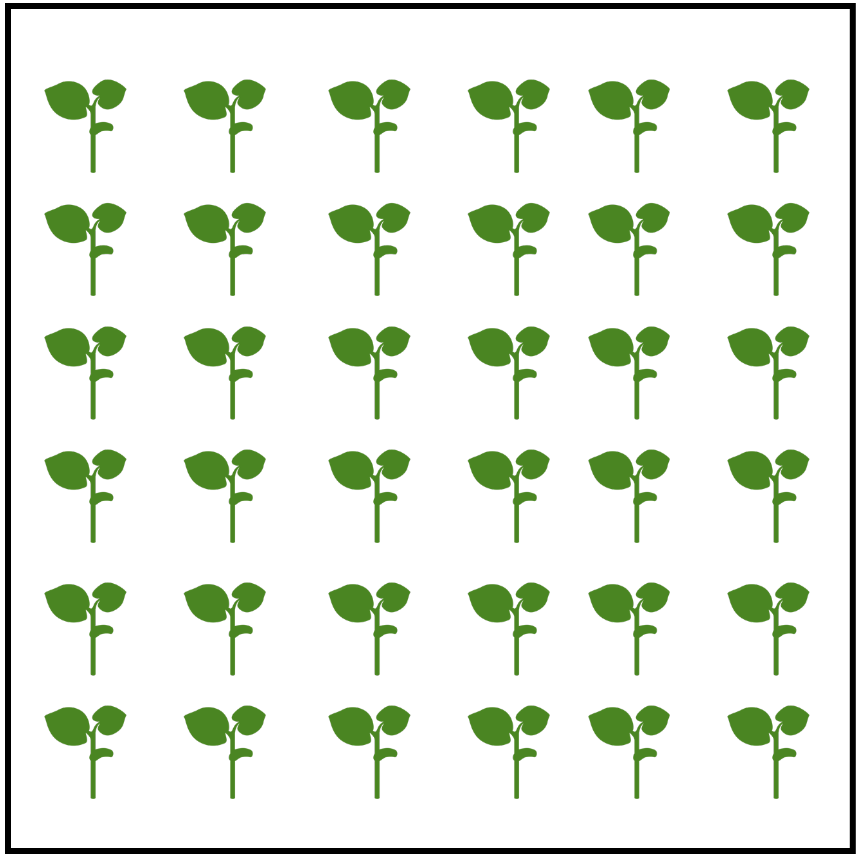
Let’s add more plants and another treatment.

Challenge plants with herbivory: inoculation of plants with 10 aphids vs. no inoculation (control).



Created by Alexandr Lavreniuk
from Noun Project

Completely randomized design (CRD; factorial)

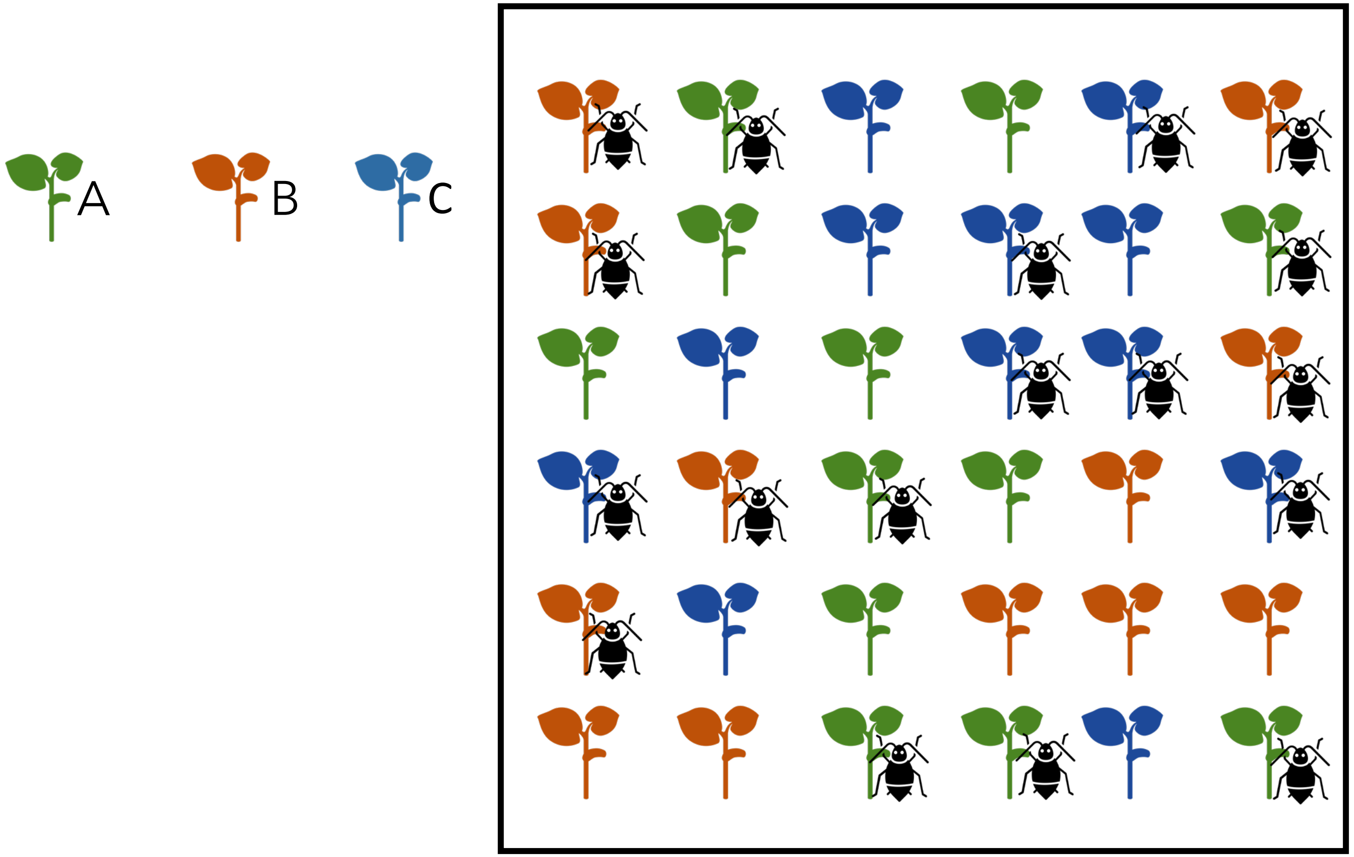


□ Sampling universe (e.g., farm)

🌱 single plant

Created by Alexandr Lavreniuk
from Noun Project

Completely randomized design (CRD; factorial)



□ Sampling universe (e.g., farm)

single plant

Created by Alexandr Lavreniuk
from Noun Project

10 aphids

Created by ProSymbols
from Noun Project

Linear model

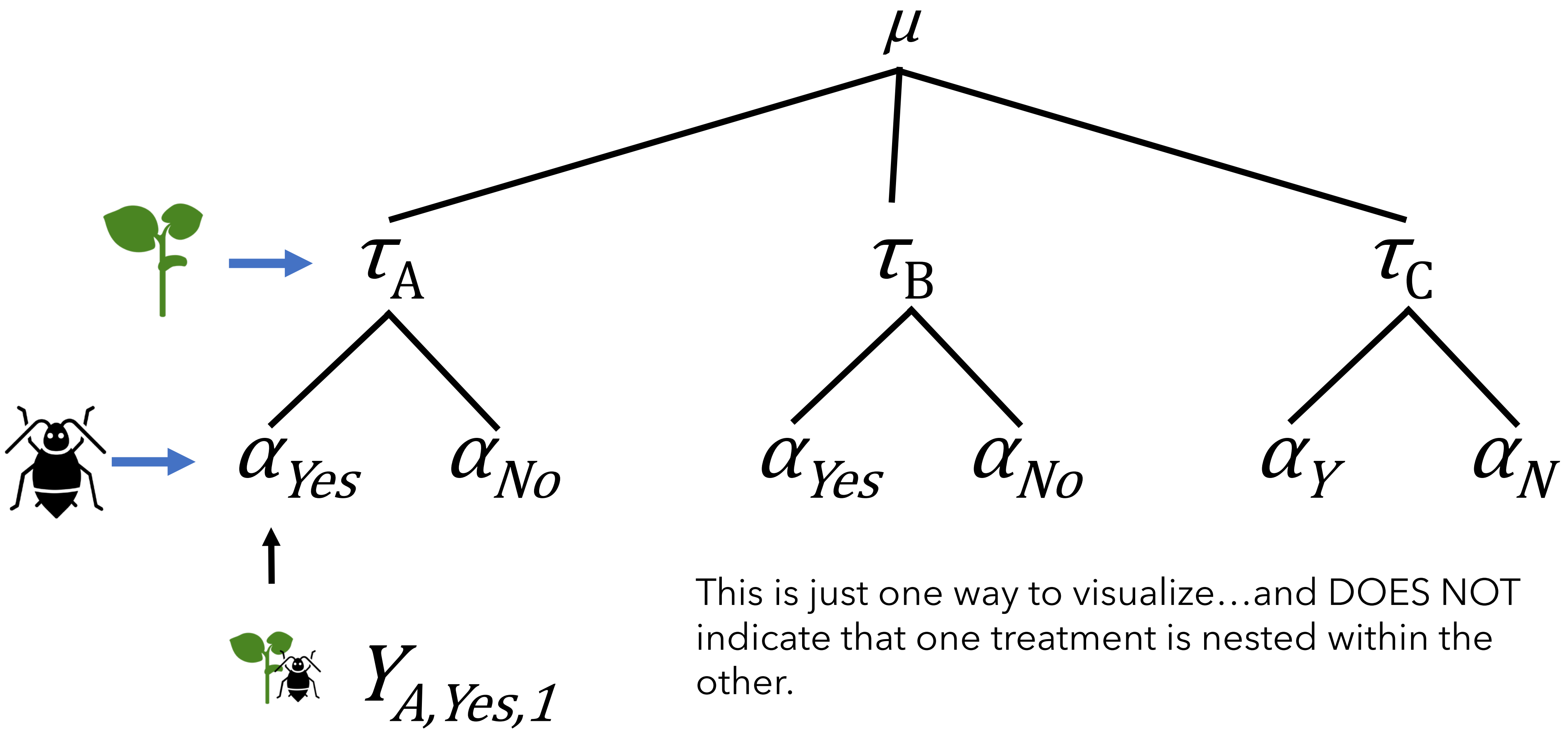
$$\underset{\text{Response}}{Y_{ijk}} = \underbrace{\mu + \tau_i + \alpha_j}_{\text{Predictors}} + \varepsilon_{ijk}$$

Y_{ij} is the value for *j*th observation in the *i*th level of the treatment.
μ is the overall mean (average) across all observations
τ_i is the effect, or the difference between the *i*th level of the treatment τ and the overall average (μ)

Now we have another treatment (aphids)…so we need another parameter:
α_j is the effect, or the difference between the jth level of the treatment α and the overall average (μ)
ε_{ijk} is the error, or the remaining difference between the *k*th observation and the mean of observations in the *i*th treatment and *j*th treatment levels

Linear model

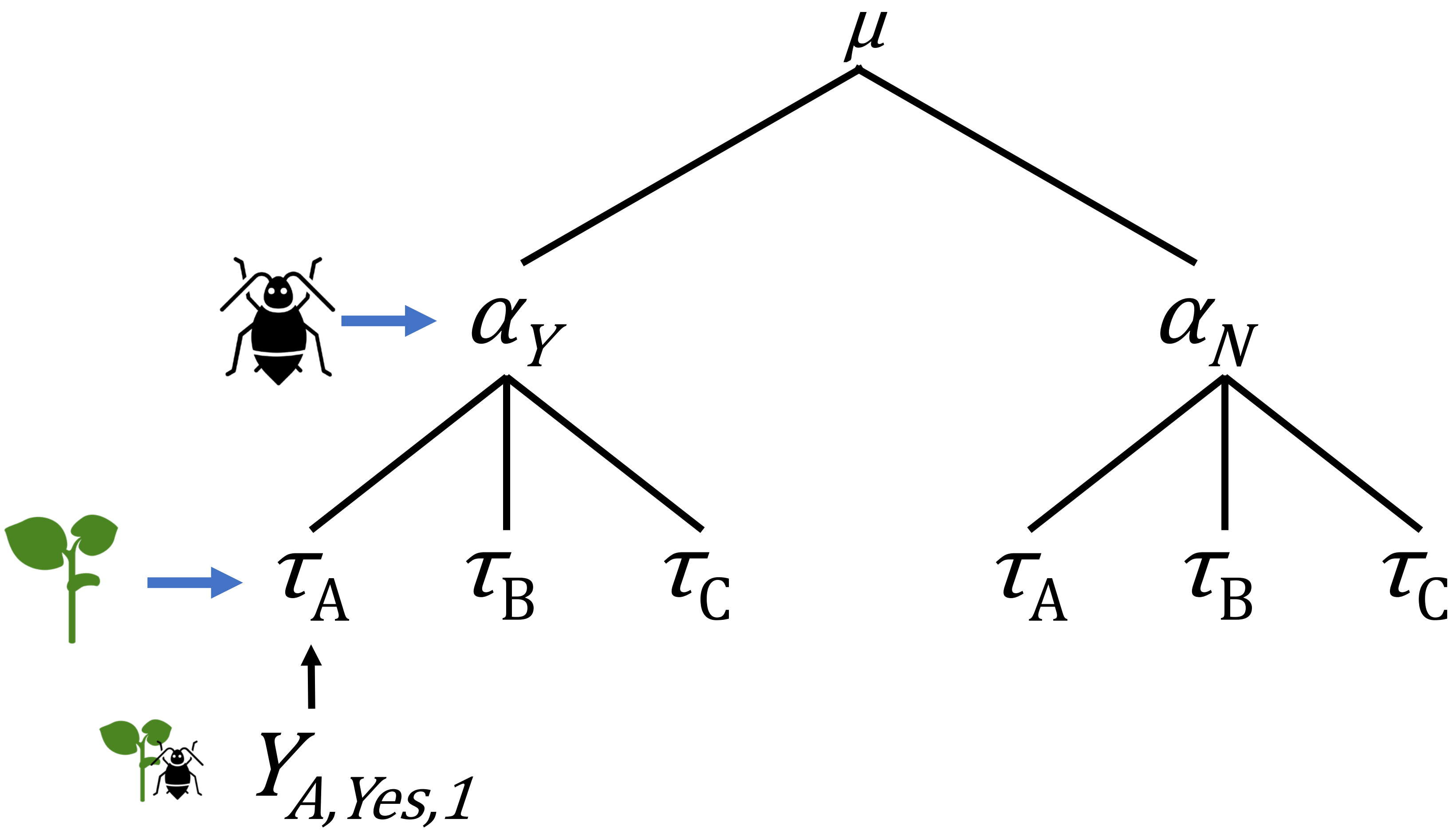
$$Y_{ijk} = \mu + \tau_i + \alpha_j + \varepsilon_{ijk}$$



This is just one way to visualize…and DOES NOT indicate that one treatment is nested within the other.

Linear model

$$Y_{ijk} = \mu + \tau_i + \alpha_j + \varepsilon_{ijk}$$



Linear model

$$\underset{\text{Response}}{Y_{ijk}} = \underbrace{\mu + \tau_i + \alpha_j + \tau\alpha_{ij}}_{\text{Predictors}} + \varepsilon_{ijk}$$

We can also evaluate interactions between our treatments (e.g., does response of plants to different fertilizer levels change when plants are challenged with aphids?)

We'll cover interactions in detail later when we cover Analysis of Variance!

Observational studies

Observational studies: when the independent or dependent variables are not manipulated by the researcher.

In observational studies, your ability to infer cause-effect relationships is limited.

Examples?

What to do if my experiment fails?

Depends on the reason for failure. It's difficult to give broadly applicable advice, but any steps you might take to rescue an experiment should be ethical (i.e., don't "torture the data until it confesses").

Some things you *might* try:

- Repeat the experiment if possible (often it's not), assuming failure was not due to a design flaw
- Combine your data with an existing data base
- Did you lose a treatment? Is there a simpler, meaningful analysis you can still try?
- Can you answer the question with a meta-analysis? (more on this in a moment…)

Null or "negative" results: Publication bias → authors are more likely to publish when they have statistically significant results ("file drawer problem")

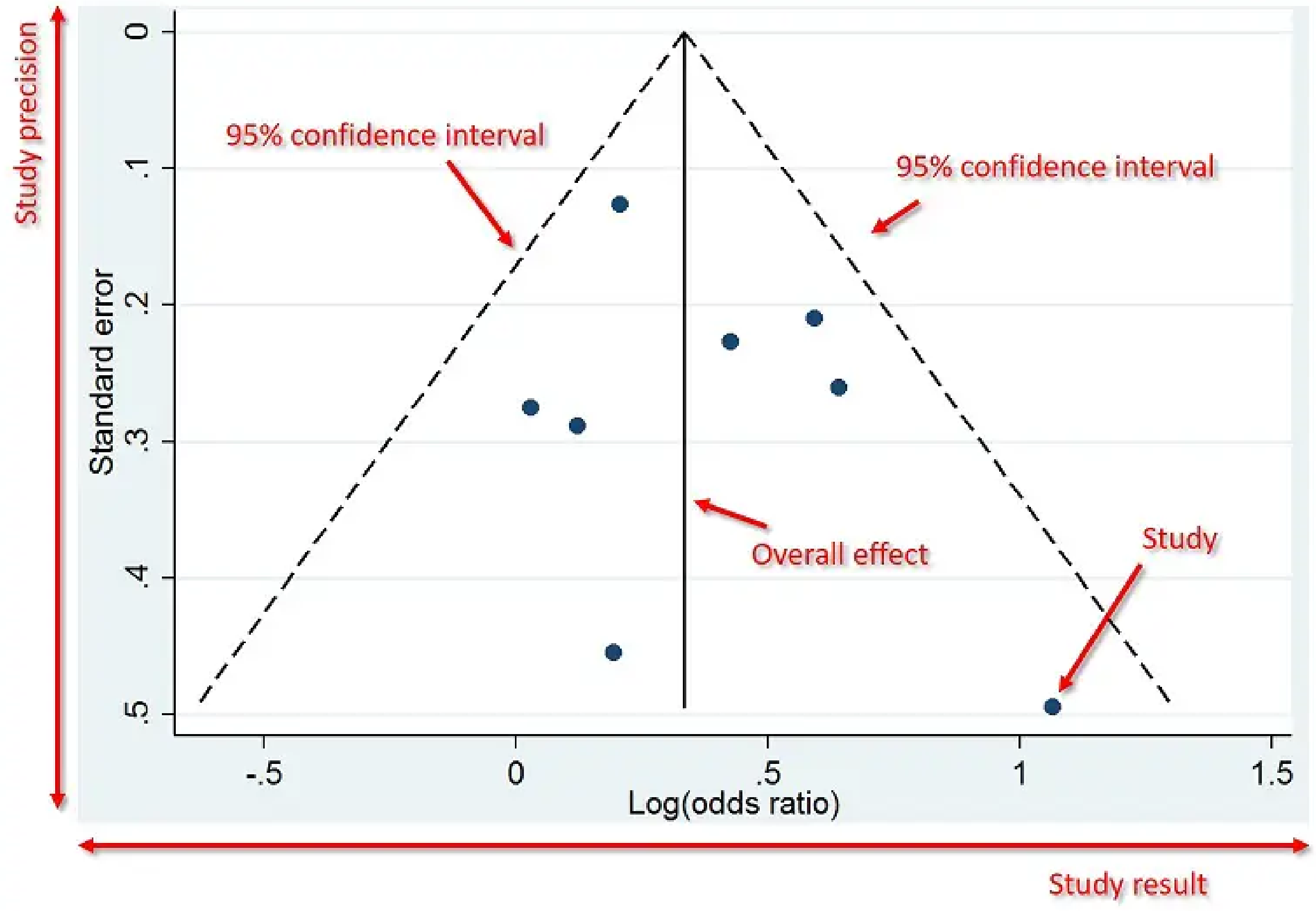
- One hallmark of good research: asking questions that are interesting no matter the answer (easier said than done, of course!).

Meta-analyses

Modern systematic reviews and meta-analyses

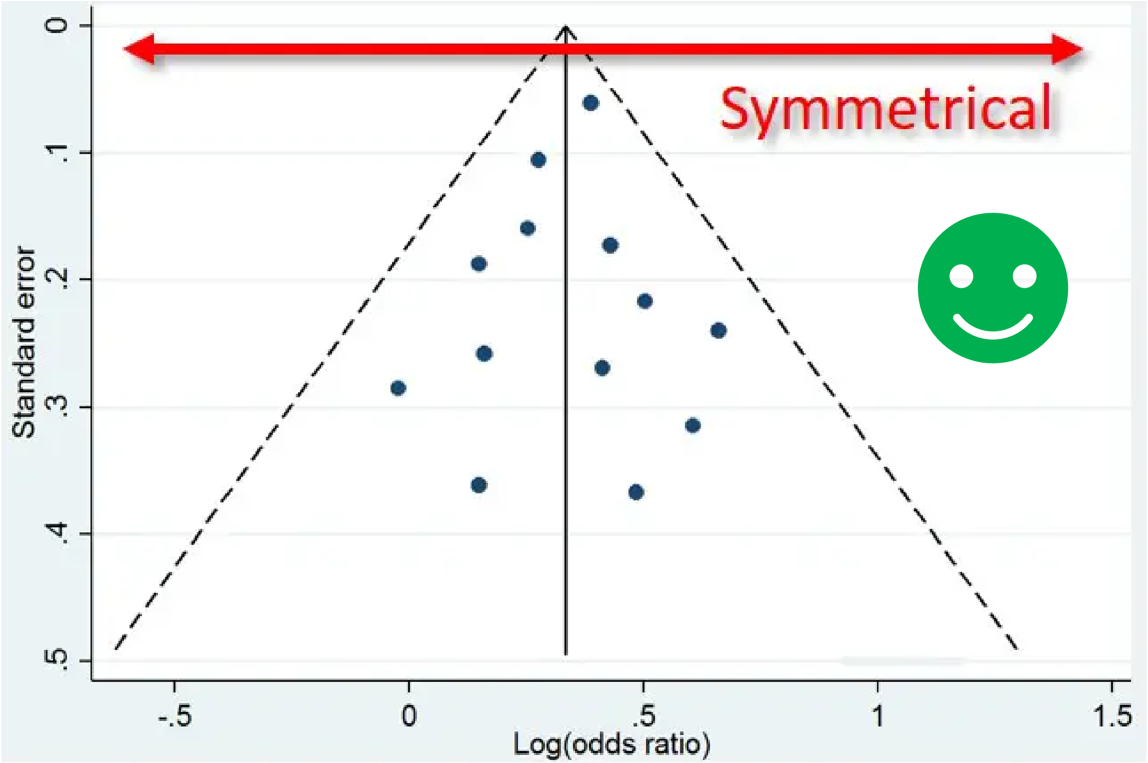
1. State question(s) to be addressed in review.
2. Annotated bibliographies vs. formal meta-analyses
3. Define in advance and report the following things
    a) methods used to locate studies
    b) list of data to be extracted, and derivation methods, if needed
    c) criteria for inclusion (and exclusion) of results based on methods used not results themselves
    d) define comparisons to be made among subgroups of studies
4. Summarize and analyze results in combined studies
    a) List studies found
    b) Explain why any were excluded
    c) Forest plots
    d) Combined, precision weighted estimate of effect(s)
    e) Contrast subgroups
    f) Sensitivity analysis ("jacknife" and other procedures)
5. Assess publication bias (file drawer effect) and methodological bias with a funnel plot

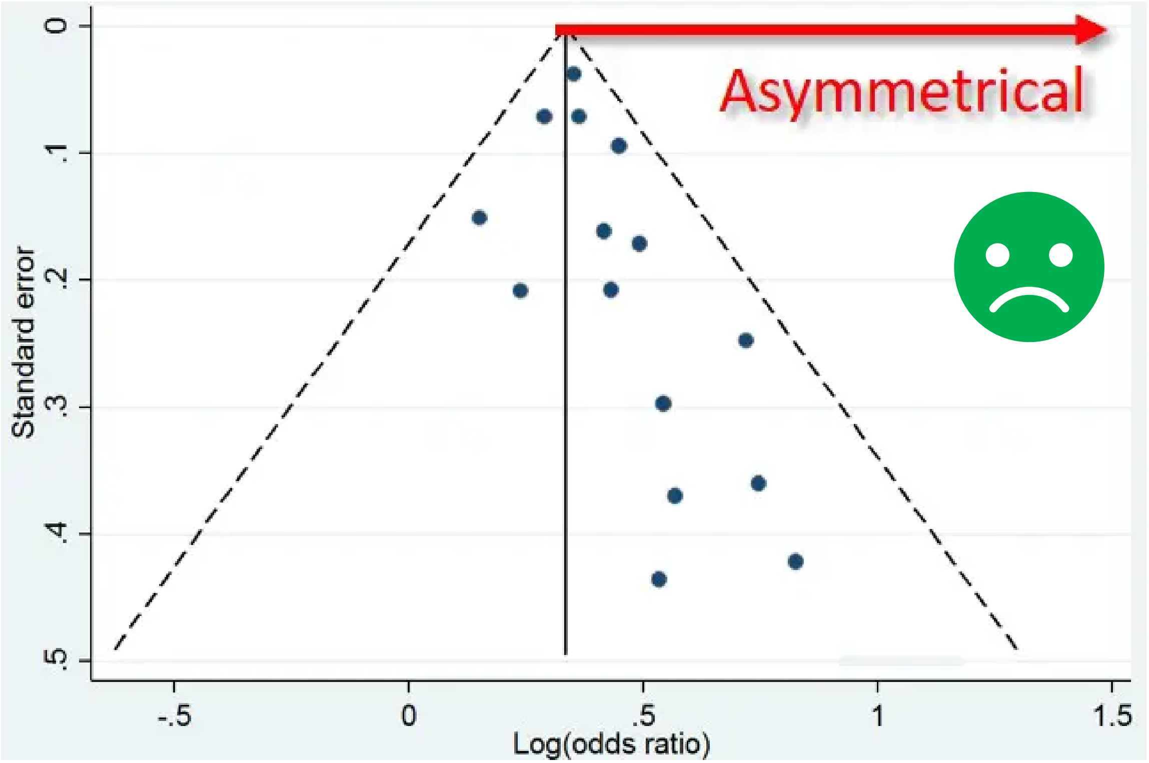
File drawer problem (funnel plots)



https://toptipbio.com/funnel-plot/

File drawer problem (funnel plots)



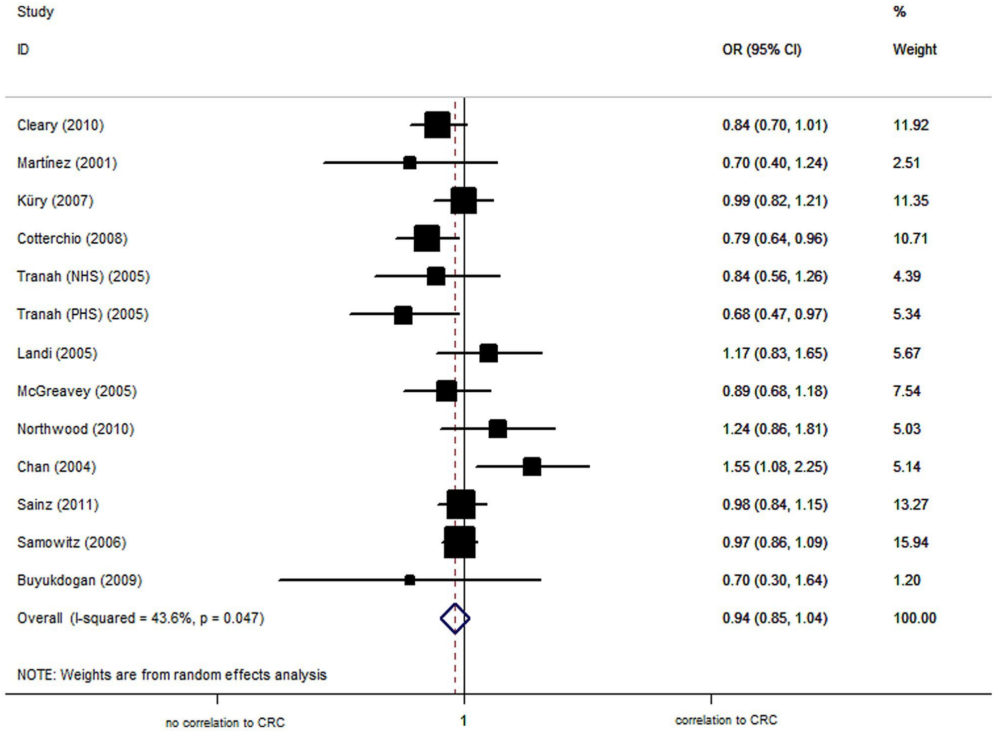


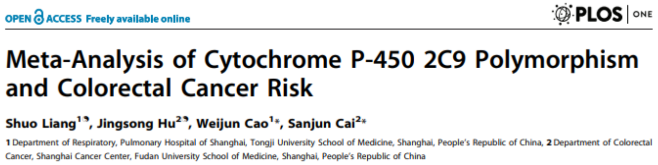
Asymmetry is indicative of publication bias

https://toptipbio.com/funnel-plot/

Forest plot example

Forest plots: graphical summary of means and variances (often 95% confidence intervals or standard errors) for each study and overall effect





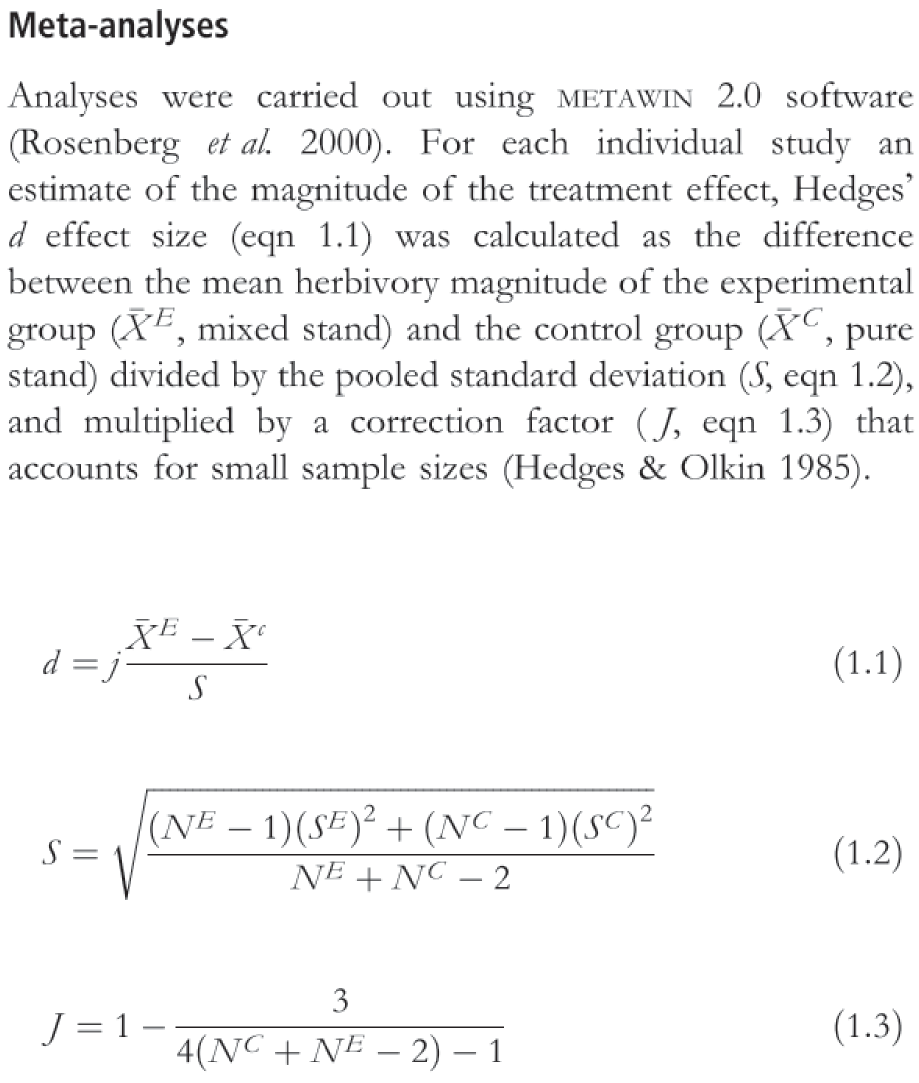
OPEN ACCESS Freely available online PLOS | ONE

Meta-Analysis of Cytochrome P-450 2C9 Polymorphism and Colorectal Cancer Risk

Shuo Liang[1¶], Jingsong Hu[2¶], Weijun Cao[1*], Sanjun Cai[2*]

1 Department of Respiratory, Pulmonary Hospital of Shanghai, Tongji University School of Medicine, Shanghai, People's Republic of China, 2 Department of Colorectal Cancer, Shanghai Cancer Center, Fudan University School of Medicine, Shanghai, People's Republic of China

Meta-analyses (example)



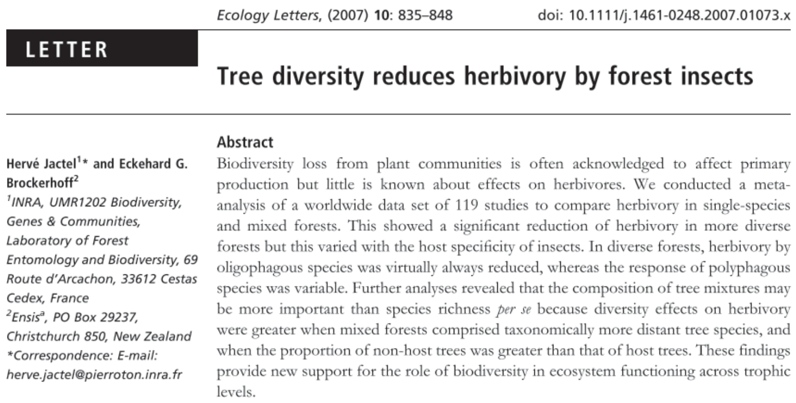
Meta-analyses

Analyses were carried out using METAWIN 2.0 software (Rosenberg *et al.* 2000). For each individual study an estimate of the magnitude of the treatment effect, Hedges' *d* effect size (eqn 1.1) was calculated as the difference between the mean herbivory magnitude of the experimental group ($\bar{X}^E$, mixed stand) and the control group ($\bar{X}^C$, pure stand) divided by the pooled standard deviation (*S*, eqn 1.2), and multiplied by a correction factor (*J*, eqn 1.3) that accounts for small sample sizes (Hedges & Olkin 1985).

$$d = j\frac{\bar{X}^E - \bar{X}^c}{S} \qquad (1.1)$$

$$S = \sqrt{\frac{(N^E - 1)(S^E)^2 + (N^C - 1)(S^C)^2}{N^E + N^C - 2}} \qquad (1.2)$$

$$J = 1 - \frac{3}{4(N^C + N^E - 2) - 1} \qquad (1.3)$$



Ecology Letters, (2007) **10**: 835–848 doi: 10.1111/j.1461-0248.2007.01073.x

LETTER

Tree diversity reduces herbivory by forest insects

Hervé Jactel[1]* and Eckehard G. Brockerhoff[2]

[1]INRA, UMR1202 Biodiversity, Genes & Communities, Laboratory of Forest Entomology and Biodiversity, 69 Route d'Arcachon, 33612 Cestas Cedex, France

[2]Ensis[a], PO Box 29237, Christchurch 850, New Zealand

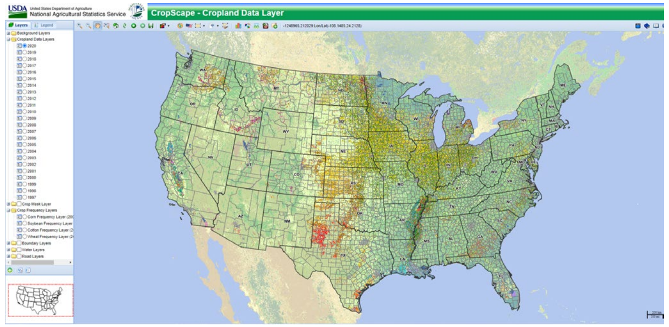
**Correspondence: E-mail: herve.jactel@pierroton.inra.fr*

Abstract

Biodiversity loss from plant communities is often acknowledged to affect primary production but little is known about effects on herbivores. We conducted a meta-analysis of a worldwide data set of 119 studies to compare herbivory in single-species and mixed forests. This showed a significant reduction of herbivory in more diverse forests but this varied with the host specificity of insects. In diverse forests, herbivory by oligophagous species was virtually always reduced, whereas the response of polyphagous species was variable. Further analyses revealed that the composition of tree mixtures may be more important than species richness *per se* because diversity effects on herbivory were greater when mixed forests comprised taxonomically more distant tree species, and when the proportion of non-host trees was greater than that of host trees. These findings provide new support for the role of biodiversity in ecosystem functioning across trophic levels.

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1461-0248.2007.01073.x

Open source data


National Agricultural Statistics Service
CropScape - Cropland Data Layer

USDA (e.g., CropScape)

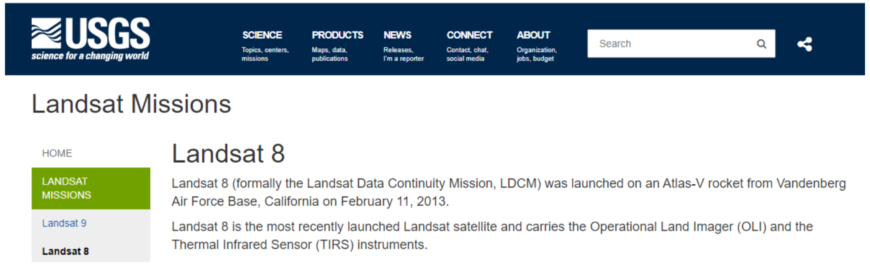
Forest Inventory and Analysis

We are the Nation's Forest Census

Welcome!

USDA Forest Service

PRISM (climate rasters)



Landsat Missions

Landsat 8

Landsat 8 (formally the Landsat Data Continuity Mission, LDCM) was launched on an Atlas-V rocket from Vandenberg Air Force Base, California on February 11, 2013.

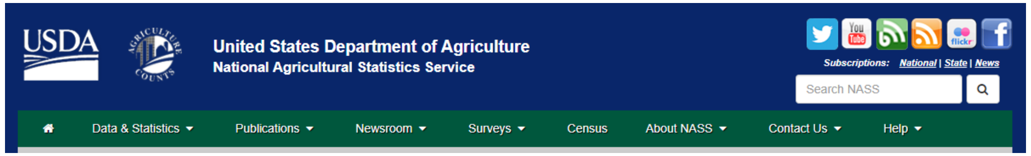
Landsat 8 is the most recently launched Landsat satellite and carries the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS) instruments.

USGS (e.g., remote sensing)

NATIONAL SCIENCE FOUNDATION

LTER NETWORK

LONG TERM ECOLOGICAL RESEARCH


USDA
United States Department of Agriculture
National Agricultural Statistics Service
Subscriptions: National | State | News
Search NASS
Data & Statistics
Publications
Newsroom
Surveys
Census
About NASS
Contact Us
Help

US Census (human population density)

NSF | neon

Operated by Battelle

NSF (National Ecological Observatory Network)