Subsampling / Variance components
  Quantify the magnitude of multiple sources of random variation

Many studies include subsampling
  Analytical Chemistry: 3 measurements on each sample
  Agronomy: 5 soil cores from a field
  Greenhouse studies: randomly assign containers to trt, measure plants
  Education: randomly assign classrooms to trt, measure students

Example: barley response to salinity (3 levels) - see plot of the data
  Randomly assign salinity level (none, low, high) to container
    2 containers per salinity level, 6 total
  Many plants per container
    randomly sample 3 plants per container, response = height after 3 weeks

Propose a 1-way ANOVA using 18 observations (6 containers x 3 plants / container)
  What are the assumptions if we use an F test?

  Which is the most important assumption?

Tools to assess the assumptions:
  Residual vs predicted value plot
  Normal quantile-quantile plot on the residuals
  Does the experimental unit match the observational unit?
    experimental unit (eu): object randomly assigned to a treatment
    observational unit (ou): object represented by one row of data

Barley salinity study:
  what is the eu?

  what is the ou?

  is there a problem with an assumption? Which?

301/587 "solution":
  average the subsamples to create one row of data for each eu

Different names for the same issue:
subsampling cluster effects biological replicates / technical replicates

Statistical approach: 2 sources of random variation:
containers
plants within a container

Nested effects: see picture drawn during lecture
Containers could be numbered 1 through 6 or 1, 2 within each treatment
If numbered 1,2 that is arbitrary
Nothing connects container 1 in the none treatment with container 1 in the low treatment
Plants could be numbered 1 through 18 or 1, 2, 3, within each container
Nothing connects plant 1 in container 1 with plant 1 in container 2
Now "higher up in the design" should make more sense.
Prefer more containers, even if same total # plants
plants are nested in containers
containers are nested in treatments (implicit when replicates of the treatment)

Models: using the barley example
Notation: i: treatment, j: container, k: plant within container
1 way ANOVA (301/587), container averages, 1 source of random variation

$$\begin{aligned} Y_{ij} &= \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim N(0,\, \sigma_e^2) \end{aligned}$$

2 sources of random variation, containers, plants within container

$$\begin{aligned} Y_{ijk} &= \mu_i + \tau_{ij} + \varepsilon_{ijk} \qquad (1) \\ \tau_{ij} &\sim N(0,\, \sigma_c^2) \quad \text{variability among containers within treatment} \\ \varepsilon_{ijk} &\sim N(0,\, \sigma_p^2) \quad \text{variability among plants within container} \end{aligned}$$

Fixed and random effects
Many different definitions / ways to distinguish
What I find most helpful: What is the inference goal?
Fixed effect, e.g. μ_i or $\mu + \alpha_i$
Goal is to estimate a treatment mean, or a regression slope
Random effect, e.g. ε_{ij}, τ_{ij}, ε_{ijk}
Goal is to estimate the variability of that random effect, i.e. σ_e^2, σ_c^2, σ_p^2
Another commonly used way to distinguish:
Effect is random if it comes from a probability distribution
Treatment means (e.g. for 3 salinity levels): 3 numbers, no randomness
Containers: many possible, not interested in mean for container # 13
Very interested in the variability between containers, σ_c^2
Plants: same argument, even more so

Very interested in the variability between plants within a container, σ_p^2
"Mixed model": a model with both
fixed effects (not counting the mean) and
random effects (not counting the error)

Two ways to quantify variability between containers:

1. Average the 3 plants within a container ⇒ one weight per container
   Compute pooled sd: container averages within treatment = 1.94
   Includes variability between plants, because averaging 3 plants
2. Imagine perfect knowledge about the container,
   compute pooled sd = 1.91
   Does not include variability between plants ("perfect knowledge")

Second quantity is called a **variance component**

Consequences of model (1):
1. observations = plants from same container are correlated, unless $\sigma_c = 0$
2. Var $Y_{ijk} = \sigma_c^2 + \sigma_p^2$
3. Var $\overline{Y}_i = \frac{\sigma_c^2}{c} + \frac{\sigma_p^2}{c\,p}$

Intraclass correlation coefficient (ICC):
Correlation between two observations in the same cluster

$$= \frac{\sigma_c^2}{\sigma_c^2 + \sigma_p^2}$$

Observations from different clusters are independent
Set of observations are not independent unless $\sigma_c^2 = 0$

Design choices: Comparing 2 treatments
You can measure 48 plants, should you
1. Use 4 containers, 12 plants per container?
2. Use 8 containers, 6 plants per container?
3. Use 24 containers, 2 plants per container?

Design choices: specifics
Want to know se of a treatment mean, se $\overline{Y}_i = \sqrt{\text{Var } \overline{Y}_i}$
need information / guesses about the variance components

Example A: $\sigma_c^2 = 4$, $\sigma_p^2 = 0.3$
1. 4 containers, 12 plants per container: Var $\overline{Y}_i = \frac{4}{4} + \frac{0.3}{4\times12} = 1.00$, se = 1.00
2. 8 containers, 6 plants per container: Var $\overline{Y}_i = \frac{4}{8} + \frac{0.3}{8\times6} = 0.51$, se = 0.71

3. 12 containers, 4 plants per container: Var $\overline{Y}_i = \frac{4}{12} + \frac{0.3}{12\times4} = 0.34$, se $= 0.58$

What if you could use twice as many plants?

1. 12 containers, 8 plants per container: Var $\overline{Y}_i = \frac{4}{12} + \frac{0.3}{12\times8} = 0.34$, se $= 0.58$
2. 24 containers, 4 plants per container: Var $\overline{Y}_i = \frac{4}{24} + \frac{0.3}{24\times4} = 0.17$, se $= 0.41$

Demonstrates what some call "hidden replication"

General principle: replicate "as high up in the design" as possible

Example B: $\sigma_c^2 = 0.1$, $\sigma_p^2 = 4.2$ (same Var $Y_{ij} = 4.3$ as before)?

1. 4 containers, 12 plants per container: Var $\overline{Y}_i = \frac{0.1}{4} + \frac{4.2}{4\times12} = 0.11$, se $= 0.34$
2. 8 containers, 6 plants per container: Var $\overline{Y}_i = \frac{0.1}{8} + \frac{4.2}{8\times6} = 0.10$, se $= 0.32$
3. 12 containers, 2 plants per container: Var $\overline{Y}_i = \frac{0.1}{12} + \frac{4.2}{12\times4} = 0.096$, se $= 0.31$

Why does Var $\overline{Y}_i$ change a lot in example A, but little in B?

Correlation between plants in the same container:

A: ICC $= 4/(4 + 0.3) = 0.93$,

multiple plants provide little new information

B: ICC $= 0.1/(4.2 + 0.1) = 0.02$,

multiple plants are essentially independent pieces of information

Applies to differences of means also.

se diff $= \sqrt{2}$ se mean, because treatments assigned to containers

split plot designs: treatments assigned to both containers and plants (to come)

Note: remember 2 ways to compute variability between containers

Previous numbers used the estimated variance components (close to example A)

Using example B information:

Variability between container averages: 1.5

Variance component for containers: 0.1

When subsamples are very variable, container averages very diff. from variance components

Return to the real barley study

Design:

3 treatments, manipulating salinity levels: none, low, high

2 containers per treatment, 3 plants per container

Data: $\Rightarrow \hat{\sigma}_c^2 = 3.64$, $\hat{\sigma}_p^2 = 0.31$

Results: Var trt mean = Var $\overline{Y}_i = \frac{3.64}{2} + \frac{0.31}{2\times3} = 1.87$

se $\overline{Y}_i = \sqrt{1.87} = 1.37$, df $= 3$

Why only 3 df? There are 18 observations.

eu = container. Care about the variability between containers

6 containers, need to estimate 3 treatment means, so df = 6-3 = 3

What if we mis-analyze the data, ignoring container

$\hat{\sigma}_e^2 = \text{MSE} = 2.49$, 15 df (= 18 - 3)
se $\overline{Y}_i = \sqrt{2.49/6} = 0.64$
CI for trt mean or difference in trt means is too narrow
no longer a 95% interval,
test now has a type I error > 5%.
details for this "misanalysis":
for this study it is a 39% CI
for this study, type I error = 61%

Estimating variance components:
Remember: σ_c^2 is the variability between containers
when have perfect knowledge about the container
Have empirical variance between container averages
Average of the three plants within a container has Var = $\sigma_c^2 + \sigma_p^2$/something
Pictures
Need to "remove" the contribution of plant-plant variability
from the variability between container averages

"Traditional" = EMS = ANOVA estimates
Calculate the average height for each container (averaging over plants)
Do a 1-way ANOVA of trt using those 6 container averages
se of a trt mean = 1.37, so variance of a trt mean = 1.87
Var $\overline{Y}_i = \frac{\sigma_c^2}{2} + \frac{\sigma_p^2}{2\times 3}$
Need an estimate of σ_p^2 = variability between plants
Do a 1-way ANOVA of container using 18 plants
$\hat{\sigma}_p^2 = \text{MSE} = 0.31$
Var $\overline{Y}_i = 1.87 = \frac{\sigma_c^2}{2} + \frac{0.31}{2\times 3}$, $\hat{\sigma}_c^2 = 3.64$
Solving for σ_c^2 involves a subtraction
Can get a negative estimate of σ_c^2

REML = restricted / residual Maximum Likelihood (ML)
ML: quantifies how well a model fits the data
more general than least squares
when errors are normally distributed, ML = Least Squares
ML estimate of a variance is biased
Solution: do ML on the residuals after fitting treatment means = REML
unbiased in simple situations, generally works better in almost all situations

Comparison of "Traditional" and REML
Traditional can give you a negative estimate of the variance component
REML, as usually implemented, will give you 0 or a positive estimate
When REML estimate > 0, usually same as Traditional
REML is now the default choice in most software
The major packages in R don't provide Traditional estimates

We'll talk more about this later

Why do anything more than 301/587 method (averaging over subsamples)?

1. When equal # plants per container
   choice doesn't matter
   301/587 or "Traditional" or REML give same answers
   so long as $\hat{\sigma}_c^2 > 0$

2. When unequal # plants per container, especially with large σ_p^2
   mixed model ("Traditional" or REML) much better
   analysis of averages assumes container averages have same variance
   Not true when # plants not constant
   $\sigma_c^2 = 0.2$, $\sigma_p^2 = 10.2$, consider Var $Y_i = \sigma_c^2 + \sigma_p^2/\#$ plants

# plants in container	Var Y_i	Result
1	0.2 / 1 + 10.2 / 1	= 10.4
2	0.2 / 1 + 10.2 / 2	= 5.3
5	0.2 / 1 + 10.2 / 5	= 2.2

   Mixed model only makes assumptions about variance components
   for containers and for plants
   gracefully handles unequal # plants

3. mixed model gives you more information
   where is the variability?
   Engineering: Gauge repeatability & reproducibility study
   $\Rightarrow$ variability between machines, between operators, between measurements
   common to report as % variability
   total variability (1 container, 1 plant) = $\sigma_c^2 + \sigma_p^2 = 3.64 + 0.31 = 3.95$
   containers = 92% of the variability, = 3.64 / 3.95
   plants = 8% of the variability, = 0.31 / 3.95

4. Have estimates of σ_c^2 and σ_p^2
   Can evaluate other choices of design
   e.g., if I measured 7 plants per container, how many containers would I need?
   se of difference between two treatment means = $\sqrt{\frac{2}{c}\left(3.64 + \frac{0.31}{7}\right)}$
   Use this se in any of the statistical sample size computations