**Stat 471/571: Analysis of repeated experiments**

Basic question: What should I do when I've done the same experiment multiple times?

In some fields, it is very difficult to publish results done at one site for one year. Two of those fields are agronomy and plant pathology. Known as repeated experiments.

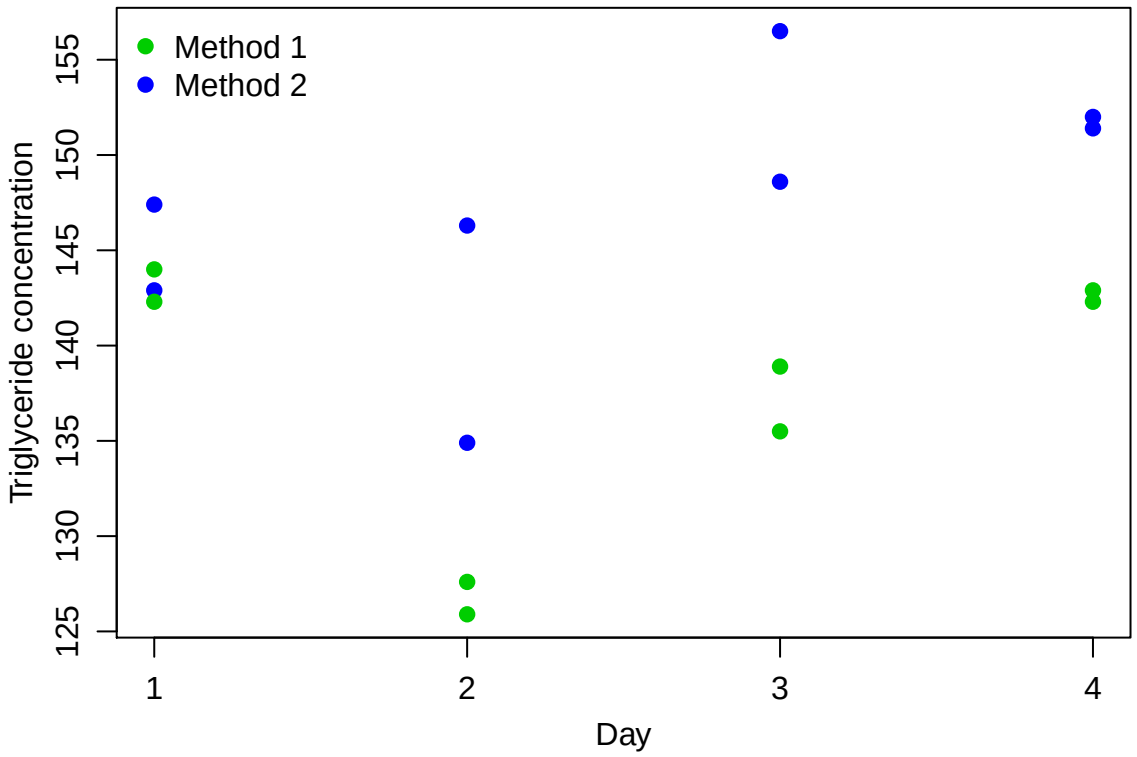
In other fields, studies are done at multiple locations for practical reasons. Example, many clinical trials, especially on rare diseases. One medical center may only see 20 cases a year. May take 10+ years to accumulate a sufficient sample size, e.g., 200 subjects. So, the study is done at 5 (or more) medical centers simultaneously. May only take 2 years to accumulate 200 subjects. Known as multi-center clinical trials.

Key difference between repeated experiments and a block design:

- Have replicates at each location / year
- so could do a separate valid analysis of each location / year
- Locations could be treated as blocks or as a non-randomized treatment
- There are more analysis options because there are replicates within each location

Example study: Triglyceride measurements in meat

- Treatments: 2 different methods to measure triglyceride concentration in a piece of meat
- Done on 4 days
- Each day, four pieces of meat randomly assigned to treatments
- So two replicates of each treatment each day
- Could do a t-test each day
  - Gives you four answers
  - Prefer one answer that includes all days




**Possible analyses**

- 1) Focus on days and methods, treat observations as subsamples
  - Calculate average for each day ⇒ one obs per method each day
  - treat days as blocks
  - | Source | df | F | p |
    |---|---|---|---|
    | Day | 3 | 3.98 | – |
    | Method | 1 | 11.25 | 0.044 |
    | Residua | 3 | | |
  - | Quantity | Estimate | se | df |
    |---|---|---|---|
    | rMSE | 4.247 | | 3 |
    | Method 1 | 148 | 2.12 | 3 |
    | Method 2 | 157 | 2.12 | 3 |
    | Difference | 10.1 | 3.00 | 3 |
- 2) Fit same model to the 16 observations
  - again treat days as blocks
  - | Source | df | F | p |
    |---|---|---|---|
    | Day | 3 | 7.08 | – |
    | Method | 1 | 19.98 | 0.0009 |
    | Residual | 11 | | |

- | Quantity | Estimate | se | df |
  |---|---|---|---|
  | rMSE | 4.51 | | 3 |
  | Method 1 | 148 | 1.59 | 11 |
  | Method 2 | 157 | 1.59 | 11 |
  | Difference | 10.1 | 2.25 | 11 |
- Where did the 11 residual df come from?
- Each day has 2 df for the error (4 obs - 1 - 1 for treatments)
- Pooling the 4 days gives 8 df for "pure error":
  variability between replicates of the same treatment on the same day
- The other 3 df are the block*treatment df

- In all previous analyses of data from block designs, we had 1 observation per block and treatment
  - Could not estimate block*treatment variability
  - When you have replicates of each block and treatment, can do that
- 3) Fit model with block*treatment interaction to the 16 observations
  - again treat days as blocks
  - but separate the "pure error" from the block*treatment interaction
  - | Source | df | F | p |
    |---|---|---|---|
    | Day | 3 | 9.97 | – |
    | Method | 1 | 28.16 | 0.0007 |
    | Method*Day | 3 | 2.50 | 0.13 |
    | Residual | 8 | | |
  - | Quantity | Estimate | se | df |
    |---|---|---|---|
    | rMSE | 3.80 | | 3 |
    | Method 1 | 148 | 1.34 | 8 |
    | Method 2 | 157 | 1.34 | 8 |
    | Difference | 10.1 | 1.90 | 8 |
- So is the se for the difference 1.90 or 3.00? Is the p-value 0.044 or 0.0007?
- Analyses 1) and 3) represent two conceptually different types of conclusions
  But the difference is subtle
- Analysis 3) answers "How precise is the difference averaged over **these four specific days**"
- Analysis 1) answers "How precise is the difference if I averaged over **four new days**"
- The difference is how the block*treatment interaction is used in the analysis.
  - In analysis 1, the error term is the B*T interaction
  - In analysis 3, the error term is the "pure error"
- Pictures to illustrate the two types of inference

**Broad-, intermediate- and narrow-sense inference**

A classic paper (1991) McLean, Sanders and Stroup, A unified approach to mixed linear models, gave names to the two types of inferences plus one more we haven't yet talked about.

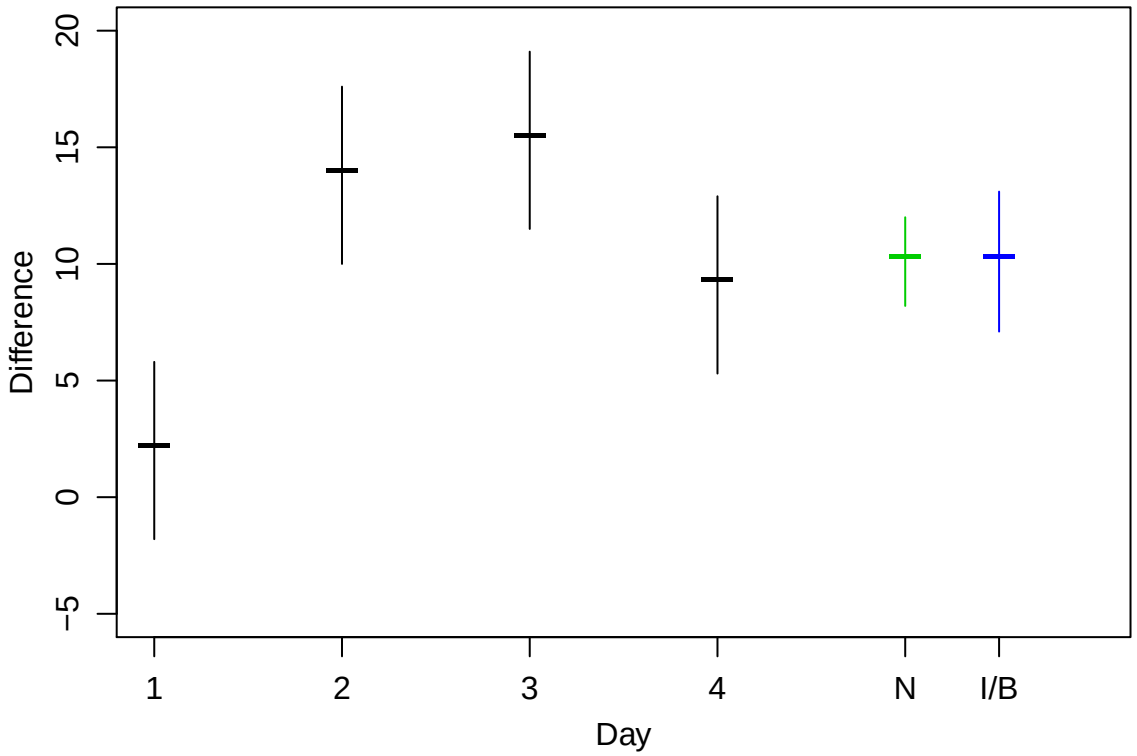
I'll use the variables in the triglyceride study, but the concepts apply equally well to locations or years.

- Narrow sense inference
  - to the specific days used in the study
  - Day*Method is a fixed effect
  - which means Day has to be a fixed effect also
  - error is the "pure error"
  - smallest standard error for the difference
  - ignores consistency or lack of consistency across repetitions
- Intermediate sense inference
  - to four new days from the same population as used in the study
  - So imagine that the four days of this study are a sample from some population of days
  - Day*Method is a random effect
  - Day is still fixed
  - error is the day*method interaction
  - results about Method depend on whether repetitions have a consistent difference or not
- Broad sense inference
  - also to four new days from the same population as used in the study
  - Both Day and Day*Method are random
  - Broad and intermediate make same inferences about Method differences
  - Different about individual means
  - Analogous to Fixed or random blocks

To summarize the relationship between types of inference and models:

| | | Narrow | | Intermediate | | Broad | |
|---|---|---|---|---|---|---|---|
| Source | df | F/R | error | F/R | error | F/R | error |
| Day | 3 | F | | F | | R | |
| Method | 1 | F | residual | F | Day*Method | F | Day*Method |
| ' Day*Method | 3 | F | | R | | R | |
| Residual | 8 | | | | | | |

Results for triglyceride study: mean *pm* se




To put the difference between narrow and broad sense inference into context

- Variance components: method*day interaction 10.8
  residual 14.4
- Which is why se for intermediate/broad inference is only ca 50% larger than narrow
- When interaction variance component >> residual, intermediate/broad is much larger

**Other potential issues: factorial treatments**

- Remember block*something interaction quantifies consistency of treatment effects across repetitions
- When multiple treatment factors, they may have different "consistencies"
- Classic example: experiments on lentils in Syria
  - 2x2x2 complete factorial: fertilizer, weevil control, weed control
  - repeated at 8 locations
  - fertilizer has relatively consistent effects at all locations
  - weevil and weed control are bit consistent

  - Author suggests that weevil and weed abundance differed between locations
  - so control had little impact at some (low abundance) and large impact at others (high abundance)
- Allow this by partitioning the block*treatment interaction into components
  - e.g. block*fertilizer, block*weevil, block*weed, etc.
- Why do I consider doing this here, and not in a standard RCBD?
  - Standard RCBD: B*T interaction is both consistency and variability among replicates
  - randomized experiment, treatments are all combinations of factors,
  - so reasonable (in my mind) to assume equal variability
  - Repeated experiment: B*T interaction is only the consistency
  - need to think about whether all factors are equally consistent
- Could apply to studies repeated over time or across locations

**Other potential issues: equal error variances**

- primarily for studies at multiple locations
- Error variance is variability between replicate plots
- This may (and probably does) differ among locations
- Can fit models (at least in SAS and R) that allow different error variances for each location
  - Some, e.g., Hans-Peter Piepho have streneously argued that unequal variances is the only legitimate analysis
- Fitting the model is much more complicated, many more numerical issues, failure to converge, etc.
- especially when relatively few error df at each location
- My experience: allowing unequal variances doesn't change conclusions very much
  - Because conclusions are about averages over locations.
  - Exception is when sample sizes are very different at each location
- But should be careful if you consider location-specific differences

**Both potential issues are examples of "to pool or not to pool"**

- Variance components (B*T, residual) are estimated by pooling variability
- What is appropriate to pool? what is not

- Data often not especially helpful
  - Can estimate a variance component from 2 observations, has 1 df,
  - Huge uncertainty (se, ci) in that estimate
  - so data doesn't give a clear picture of similar or not
- My advice: consider the data and also your knowledge of its context
- I tend to pool unless there are reasons otherwise
- Others are splitters: don't pool unless reasons to do so