STAT 5200 Handout #5.a CRD Tests & ANOVA Table (Ch. 3)

Recall beet lice example of Handout #4: four treatments (A, B, C, D)
→ more than just two-sample t-test; this example is actually a Completely Randomized Design

CRD characteristics:

factor has multiple levels (1, 2, ..., g)
- each level corresponds to a specific "treatment"

1. a single experimental factor with $g \geq 2$ levels
   ↳ # levels ($g = 4$ here)
2. experimental units are randomly assigned to factor levels
   ⇒ all arrangements are equally likely
3. one measurement of response variable is made on each experimental unit
   - if multiple response variables are of interest, we analyze each one separately
4. if # of experimental units assigned to each factor level is the same (say n), then we say this design is balanced
   ↳ here, $n = 25$

↳ in this course, we're focusing on balanced designs because they tend to have more statistical power

Two main ways to analyze a CRD:

I. Means model

II. Effects model

I. Means model

"Y sub i j"

example of a "parameter" – a fixed, unknown value

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

Y_{ij}: Value of response variable for j^{th} experimental unit in factor level i

μ_i: mean for factor level i (in population)

ϵ_{ij}: residual error or random error "difference"

$i = 1, \ldots, g$ $\quad$ $j = 1, \ldots, n$ (or n_i if unbalanced) $\quad$ $N = g \cdot n$ (overall sample size)

Test of primary interest: $H_0 : \mu_1 = \ldots = \mu_g$

H_a: μ_i's not all equal

Can estimate pop. means μ_i with: sample means

$$\hat{\mu}_i = \bar{Y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij} = \text{sample mean in factor level } i$$

"Y i dot bar"

Construct test statistic to measure size of differences among $\hat{\mu}_i$'s

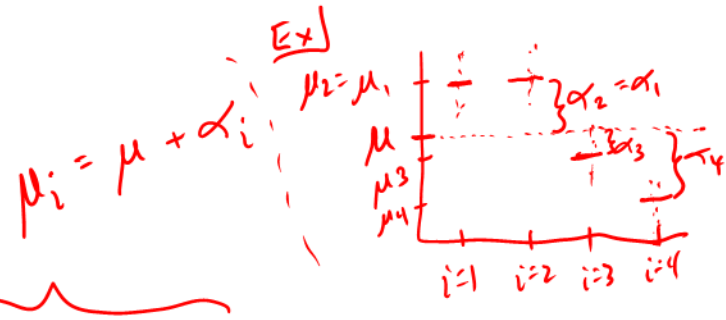
↳ will be a standardized difference

Means model discussion points:

- very simple & intuitive
- easy to formulate sensible hypotheses

- obvious parameter estimates for μ_i's, with nice mathematical properties
- BUT – hard to generalize to more than one factor

– especially when combined effect of multiple factors is of interest



II. Effects model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- Y_{ij}: value of response variable for j^{th} exp. unit in factor level i
- μ: overall or grand mean
- α_i: effect of factor level i
- ϵ_{ij}: residual error

$i = 1, \ldots, g$ $\quad$ $j = 1, \ldots, n$ $\quad$ $N = g \cdot n$

Look at # of parameters:

- Here, there are $\mu, \alpha_1, \alpha_2, \ldots, \alpha_g \rightarrow g+1$
- But how many groups of data? g (# factor levels)

Need to impose some "identifiability" constraint to ensure unique estimates

- Default in SAS is to constrain: $\alpha_g \equiv 0$, so really only have g parameters
  - Then $\hat{\mu} = \bar{Y}_{g\cdot}$, and for each $i = 1, \ldots, g-1$, $\quad \hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{g\cdot}$
  - So for every i, $\quad \hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i = \bar{Y}_{i\cdot}$
- Another option (appropriate for balanced designs, and often useful for pedagogical purposes) is the "sum-to-zero" constraint:

$$\sum_{i=1}^{g} \alpha_i = 0$$

$$\hat{\mu} = \bar{Y}_{\cdot\cdot} = \frac{1}{N}\sum_{ij} Y_{ij} = \frac{1}{g \cdot n}\sum_{i=1}^{g}\sum_{j=1}^{n} Y_{ij} = \frac{1}{g}\sum_{i=1}^{n} \bar{Y}_{i\cdot}$$

$$\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$$

$$\sum_{i=1}^{g} \hat{\alpha}_i = \sum_{i=1}^{g} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}) = \sum_{i=1}^{g} \bar{Y}_{i\cdot} - g\bar{Y}_{\cdot\cdot} = 0$$

Compare the means model and effects model:

$$\begin{aligned} Y_{ij} &= \mu_i + \epsilon_{ij} \\ &= \mu + \alpha_i + \epsilon_{ij}, \\ &\quad (i = 1, \ldots, g \; ; \; j = 1, \ldots, n) \end{aligned}$$

means model
effects model

$\hat{Y}_{ij} = \hat{\mu}_i$
$= \hat{\mu} + \hat{\alpha}_i$

In both models, parameter estimates are obtained by minimizing: $\sum_{ij} \left(Y_{ij} - \hat{Y}_{ij}\right)^2$

make predicted Y ($\hat{Y}$) close to observed Y

$\hat{\epsilon}_{ij} = e_{ij}$ { residual for exp. unit j in factor level i

- this is the principle of least squares, and gives estimates nice properties

minimum bias
& minimum variance

To construct a test statistic for $H_0 : \mu_1 = \ldots = \mu_g$ (or equivalently, $H_0 : \alpha_1 = \ldots = \alpha_g$), we need to "decompose" observations into statistics. Here, we'll focus on the effects model, with sum-to-zero constraint):

Y_{ij}

$\frac{1}{n}\sum_{j=1}^{n} Y_{ij} = \bar{Y}_{i\cdot}$

$\frac{1}{g}\sum_{i=1}^{g} \bar{Y}_{i\cdot} = \bar{Y}_{\cdot\cdot}$

$$\begin{aligned} Y_{ij} &= \mu + \alpha_i + \epsilon_{ij} \\ Y_{ij} &= \bar{Y}_{\cdot\cdot} + \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right) + \left(Y_{ij} - \bar{Y}_{i\cdot}\right) \end{aligned}$$

$\hat{Y}_{ij} = \bar{Y}_{i\cdot}$

$= a + b + c$

Square both sides and sum over both subscripts: $(a + b + c)^2 = a^2 + b^2 + c^2 + \ldots$

cross-product terms

$$\begin{aligned} \sum_{i=1}^{g}\sum_{j=1}^{n} Y_{ij}^2 &= \sum_{i=1}^{g}\sum_{j=1}^{n} \left(\bar{Y}_{\cdot\cdot}\right)^2 + \sum_{i=1}^{g}\sum_{j=1}^{n} \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 + \sum_{i=1}^{g}\sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 \\ &\quad + 2\sum_{ij} \left[\bar{Y}_{\cdot\cdot}\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right) + \bar{Y}_{\cdot\cdot}\left(Y_{ij} - \bar{Y}_{i\cdot}\right) + \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)\left(Y_{ij} - \bar{Y}_{i\cdot}\right)\right] \end{aligned}$$

these cross-product terms cancel out & sum to 0

$$= N\bar{Y}_{\cdot\cdot}^2 + n\sum_{i=1}^{g} \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 + \sum_{i=1}^{g}\sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$$

Now subtract $N\bar{Y}_{\cdot\cdot}^2$ from both sides:

$\hat{\alpha}_i$

$e_{ij} = Y_{ij} - \hat{Y}_{ij}$

$$\sum_{i=1}^{g}\sum_{j=1}^{n} Y_{ij}^2 - N\bar{Y}_{\cdot\cdot}^2 = n\sum_{i=1}^{g} \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 + \sum_{i=1}^{g}\sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$$

Sum of squares / total $\rightarrow$ $SS_T \propto \text{Var}(Y)$ SS_{Trt} SS_E

overall sample mean

SS_T = corrected total sum of squares:

$$
\begin{aligned}
(N-1)Var(Y) &= \sum_{ij}\left(Y_{ij}-\bar{Y}_{\cdot\cdot}\right)^2 \\
&= \sum_{ij}\left(Y_{ij}^2-2Y_{ij}\bar{Y}_{\cdot\cdot}+\bar{Y}_{\cdot\cdot}^2\right) \\
&= \sum_{ij}Y_{ij}^2-2\bar{Y}_{\cdot\cdot}\sum_{ij}Y_{ij}+N\bar{Y}_{\cdot\cdot}^2 \\
&= \sum_{ij}Y_{ij}^2-2\bar{Y}_{\cdot\cdot}\left(N\bar{Y}_{\cdot\cdot}\right)+N\bar{Y}_{\cdot\cdot}^2 \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n}Y_{ij}^2-N\bar{Y}_{\cdot\cdot}^2 \quad = SS_T \propto Var(Y)
\end{aligned}
$$

SS_{Trt} = sum of squares due to treatment (factors in model)

↳ like variance between factor levels

SS_E = sum of squares due to error; also "residual sum of squares": $\sum_{ij} e_{ij}^2 = \sum_{ij}\left(Y_{ij}-\hat{Y}_{ij}\right)^2$, $\hat{Y}_{ij} = \bar{Y}_{i\cdot}$

(This is the quantity that our parameter estimates are guaranteed to minimize.)

↳ like variance within factor levels

Note that $SS_{Trt} = n\sum_{i=1}^{g}\hat{\alpha}_i^2$, so we can use this to test

"no Trt. effect" → $H_0 : \alpha_1 = \ldots = \alpha_g = 0$ (implied by constraint) (or $\mu_1 = \ldots = \mu_g$)

vs.

$H_a : \alpha_i \neq 0$ for at least one i (or $\mu_i \neq \mu_{i'}$ for some i, i')

- If H_0 is true, SS_{Trt} should be: smaller
- If H_a is true, SS_{Trt} should be: larger
- How large must SS_{Trt} be to say H_a is true? Need to standardize first

Test statistic for $H_0 : \alpha_1 = \ldots = \alpha_g$:

$$F_{Trt} = \frac{SS_{Trt}/(g-1)}{SS_E/(N-g)} = \frac{MS_{Trt}}{MS_E}$$

Statistical theory says under H_0, $F_{Trt} \sim F_{g-1,N-g}$ (sampling dist'n)

under H_0: if H_0 true and model assumptions are met (HW#6 & 6a)

Sampling distribution is F with degrees of freedom: $g-1$ numerator, $N-g$ denominator

→ area = P-value = $P\{F_{g-1,N-g} > F_{Trt}\}$

We summarize this decomposition and test statistic in the ANOVA table:

- "Analysis of Variance": we partition (or decompose) variance (in response Y) into components, or sources (of variance)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	$g-1$	SS_{Trt}	$MS_{Trt} = \frac{SS_{Trt}}{g-1}$	$F_{Trt} = \frac{MS_{Trt}}{MS_E}$	P-value $= P\{F_{g-1,N-g} > F_{Trt}\}$
Error	$N-g$	SS_E	$MS_E = \frac{SSE}{N-g}$		
Corrected Total	$N-1$	SS_T			

In beet lice data (Handout # 5), $F_{Trt} =$ 13.86 , with p-value < .0001

Conclude: At least two chemical treatments have different effects

Now – which two (or more) factor levels are significantly different from each other?

- This involves post hoc mean comparisons (Ho # 7 & 7a)
- We'll come back to this later, because before we can trust any inference, we need to assess model assumptions (Ho #6 & 6a)

NOTE: DF = degrees of freedom, or # of unconstrained data points

- Every time you estimate a parameter, you lose one DF
- With sum-to-zero constraint: N = total sample size → $\sum_i \alpha_i = 0$
  - Estimate $\hat{\mu} = \bar{Y}_{..} \rightarrow$ lose 1 $DF \rightarrow DF$ for SS_T is $N-1$
    (Knowing average $\bar{Y}_{..}$ and Y_{ij} values for $N-1$ observations would allow you to calculate the remaining observation)
  - Estimate $\hat{\alpha}_1, \ldots, \hat{\alpha}_{g-1} \rightarrow DF$ lose $g-1$ $DF \rightarrow DF$ for SS_{Trt} is $g-1$
    (Why not α_g?) → constraint
  - This leaves $N-g$ DF for SS_E
    → lost g DF to get there

Another useful statistic from ANOVA table:

- $SS_T = SS_{Trt} + SS_E$

  SS_E ∝ variation in y unexplained by model

  SS_{Trt} ∝ variation in y "explained" by model

  SS_T ∝ total variation in y

$$R^2 = \frac{SS_{Trt}}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

$= $ proportion of variation (or variability) in observed response values "explained" by differences in factor level effects (trt. effects)

- $0 \leq R^2 \leq 1$
  - small R^2: variability not really explained by factor levels
  - large R^2: factor levels cause most of the variability

- If $R^2 \approx 1$, then even small differences among factor levels may be "statistically significant" (even if inconsequential)
- How high must R^2 be for a "good" model?

  - don't rely too much on this; it's usually best to just report R^2

Aside:

linear regression also has R^2 where interpretation is same and R is corr(X, Y)