Estimation?

- Talked exclusively about hypothesis testing
- What about point estimates of:
  - means?
  - treatment effects and contrasts?
  - variance components?
- What about confidence intervals?
- What about correlations between observations?

Recap

Example 9.3

- Two-factor, random effects, factorial design
- Inference:
    1. Overall F test
    2. Individual F tests, if necessary
    3. Estimate variance components (point! interval?)
    4. Estimate mean response (point! interval!)
- Plan:
    1. Fit model
    2. Check assumptions
    3. Determine correct tests
    4. Carry out inference procedures
    5. Make appropriate contextual conclusions

Note: Original data did not meet normality assumption. log transformed data is used instead.

F Tests!

Example 9.3

- **Overall:** $H_0 : \sigma^2_{Sample} = \sigma^2_{Lab} = \sigma^2_{Sample*Lab} = 0$ vs. H_1 : at least one is positive
  $F = 191.44$ and $p < 0.0001$
- **Interaction:** $H_0 : \sigma^2_{Sample*Lab} = 0$ vs. $H_1 : \sigma^2_{Sample*Lab} > 0$
  $F = 2.94$ and $p = 0.0161$
- **Sample Effect:** $H_0 : \sigma^2_{Sample} = 0$ vs. $H_1 : \sigma^2_{Sample} > 0$
  $F = 391.94$ and $p < 0.0001$
- **Lab Effect:** $H_0 : \sigma^2_{Lab} = 0$ vs. $H_1 : \sigma^2_{Lab} > 0$
  $F = 12.72$ and $p = 0.0003$
- **Conclusion:** There is evidence that in addition to the variability present between labs and samples a significant amount of variability exists due to the Sample×Lab interaction. It appears that the interlaboratory effects vary by sample - i.e. the differences between labs is not constant across different samples!

Checking our Math

- Compare these results to the tests provided by SAS - why don't our results match!
- How did I know SAS was wrong and how do we fix it!?
  - In SAS $F_{Sample} = \frac{MS\ Sample}{MSE}$
  - What do the EMS tell us the test ***should*** be?
  - $F_{Sample} = \underline{\hspace{3cm}}$
- Need to add the TEST statement into our PROC GLM

  Or, you know, actually use PROC MIXED ...

Estimating Variance Components

Example 9.3

- From our SAS output we know $MSA = 17.7299$, $MSB = 0.5756$, $MSAB = 0.0452$, and $MSE = 0.0154$.
- Estimating via our Type III EMS (A = Sample, B = Lab) we have

$$
\begin{aligned}
MSA &= \hat{\sigma}^2 + 2\hat{\sigma}^2_{AB} + 10\sigma^2_A \\
MSB &= \hat{\sigma}^2 + 2\hat{\sigma}^2_{AB} + 8\sigma^2_B \\
MSAB &= \hat{\sigma}^2 + 2\hat{\sigma}^2_{AB} \\
MSE &= \hat{\sigma}^2
\end{aligned}
$$

- Substitution yields

$$
\begin{array}{cccccc}
\hat{\sigma}^2 & = & MSE & = & & = & 0.0154 \\
\hat{\sigma}^2_{AB} & = & \frac{MSAB - MSE}{n} & = & \frac{0.0452-0.0154}{2} & = & 0.0149 \\
\hat{\sigma}^2_{A} & = & \frac{MSA - MSAB}{nb} & = & \frac{0.5756-0.0452}{8} & = & 0.0663 \\
\hat{\sigma}^2_{B} & = & \frac{MSB - MSAB}{na} & = & \frac{17.7299-0.0452}{10} & = & 1.7680
\end{array}
$$

Or just use PROC MIXED!

Estimating the Mean Response

Example 9.3

- $\hat{\mu} = \overline{y}_{...} = 6.8156$ (on the log scale)
- What about a standard error?

$$V\left[\overline{y}_{...}\right] = \frac{\sigma_A^2}{a} + \frac{\sigma_B^2}{b} + \frac{\sigma_{AB}^2}{ab} + \frac{\sigma^2}{abn}$$

- How do we estimate that!?

$$\begin{aligned} \widehat{SE}\left[\overline{y}_{...}\right] &= \sqrt{\frac{\hat{\sigma}_A^2}{a} + \frac{\hat{\sigma}_B^2}{b} + \frac{\hat{\sigma}_{AB}^2}{ab} + \frac{\hat{\sigma}^2}{abn}} \\ &= \text{lots of algebra and cancelling} \\ &= \sqrt{\frac{1}{abn}\left(MSA + MSB - MSAB\right)} \end{aligned}$$

- For the Milk Pasteurization Example this becomes

$$\widehat{SE}\left[\overline{y}_{...}\right] = \sqrt{\frac{1}{40}\left(17.7299 + 0.5756 - 0.0452\right)} = 0.6757$$

Confidence Interval for the Mean Response

- Easy! $\bar{y}_{...} \pm \left(t_{\alpha/2,df}\right)\left(\widehat{SE}\right)$
- Degrees of freedom? We need Satterthwaite's formula again - but more general.
- Our standard error is of the form $\sqrt{\sum_{i=1}^{k} c_i MS_i}$
  (A linear combination of Mean Squares)
- Satterthwaite's general formula for df in this case is

$$\widehat{df} = \frac{\left(\sum_{i=1}^{k} c_i MS_i\right)^2}{\sum_{i=1}^{k} \frac{(c_i MS_i)^2}{df_i}}$$
$$= \frac{(c_1 MS_1 + c_2 MS_2 + \cdots + c_k MS_k)^2}{\frac{(c_1 MS_1)^2}{df_1} + \frac{(c_2 MS_2)^2}{df_2} + \cdots + \frac{(c_k MS_k)^2}{df_k}}$$

CI for the Mean - Finally!

Example 9.3

- Easy? $\bar{y}_{...} \pm \left(t_{\alpha/2,df}\right)\left(\widehat{SE}\right)$
- Degrees of freedom?

$$\widehat{df} = \frac{(0.6757^2)^2}{\frac{(17.73^2/40)^2}{3} + \frac{(0.5756^2/40/)^2}{4} + \frac{(-0.0452^2/40)^2}{12}} = 3.18$$

- Log-scale interval?

$$\begin{aligned}\bar{y}_{...} \pm \left(t_{\alpha/2,df}\right)\left(\widehat{SE}\right) &= 6.52 \pm \left(t_{0.025,3.18}\right)(0.6757)\\ &= 6.8156 \pm 3.08\,(0.6757)\\ &= 6.8156 \pm 2.08\\ &= (4.7356, 8.8956)\end{aligned}$$

Summary: Milk Pasteurization Example

- Lab-to-lab variability depends on the sample they measure
  (That's bad...)
- Overall mean log-count is between 4.74 and 8.90 with 95% confidence.
- Equivalently, mean count is between 114.43 and 7331.97 with 95% confidence.
- Random effects model requires special care in software: GLM requires specifying correct EMS; need to ensure correct df are used.

Recap

Example 9.4

- Two-factor, mixed effects, factorial design
- Inference:
  1. Individual F tests, if necessary
  2. Contrasts for fixed effects, if necessary
  3. Estimate variance components
  4. Estimate mean response
  5. Estimate response correlations (New!)
- Plan:
  1. Fit model
  2. Check assumptions
  3. Determine correct tests
  4. Carry out inference procedures
  5. Make appropriate contextual conclusions

Note: Original data did not meet normality assumption. log transformed data is used instead.

F Tests!

Example 9.4

- **Interaction:** $H_0 : \sigma^2_{Day*Location} = 0$ vs. $H_1 : \sigma^2_{Day*Location} > 0$

  $F = 1.38$ and $p = 0.2303$

- **Location Effect:** $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ vs. H_1 : at least one is non-zero

  $F = 43.17$ and $p = 0.0002$

- **Day Effect:** $H_0 : \sigma^2_{Day} = 0$ vs. $H_1 : \sigma^2_{Day} > 0$

  $F = 1.84$ and $p = 0.2375$

- **Conclusion:** There is insufficient evidence that the day has a significant on the overall variance of log *Campylobacter* counts - either through an interaction with location or on its own. (I.e. not only is there no significant evidence that the day-to-day variability differs across locations, there's no evidence of a significant day-to-day variance contribution.) However, there is sufficient evidence to suggest that not all locations have the same effect on the mean log *Campylobacter* count.

Contrasts?

- PROC MIXED results layout is different - *i* and *j* replaced by *EFFECT* and *_EFFECT*
- PROC MIXED does not produce a profile plot, but could use GLM
- Output also shows original and adjusted columns
- Results suggest Locations 1 and 2 aren't different and Locations 3 and 4 aren't different, but all other locations are different at the overall 5% significance level.

Estimating Variance Components

Example 9.4

- From our SAS output we know $MSA = 32.6218$, $MSB = 1.3937$, $MSAB = 0.7556$, and $MSE = 0.5487$.
- Estimating via our Type III EMS (A = Location, B = Day) we have

$$
\begin{aligned}
MSA &= \hat{\sigma}^2 + 30\psi_A^2 + 10\sigma_{AB}^2 \\
MSB &= \hat{\sigma}^2 + 40\hat{\sigma}_B^2 + 10\sigma_{AB}^2 \\
MSAB &= \hat{\sigma}^2 + 10\hat{\sigma}_{AB}^2 \\
MSE &= \hat{\sigma}^2
\end{aligned}
$$

- Substitution yields

$$
\begin{array}{ccccccc}
\hat{\sigma}^2 & = & MSE & = & & = & 0.5487 \\
\hat{\sigma}_{AB}^2 & = & \frac{MSAB-MSE}{n} & = & \frac{0.7556-0.5487}{10} & = & 0.0207 \\
\hat{\sigma}_B^2 & = & \frac{MSB-MSAB}{na} & = & \frac{1.3937-0.7556}{40} & = & 0.0160
\end{array}
$$

Or just use PROC MIXED!

Implied Correlations

- Responses - any two y values - used to be independent
- Introducing random factors results in dependent responses
- For this example, what are the correlations of two observations taken:
  - at the same location, on the same day?
  - at different locations, on the same day?
  - on different days?

Back to ST 511!

- Recall that $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$
- Also recall that
    - $\beta_i \overset{iid}{\sim} N(0, \sigma^2_B)$
    - $(\alpha\beta)_{ij} \overset{iid}{\sim} N(0, \sigma^2_{AB})$
    - $\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$
    - and that each random effect is independent of other random effects
- Recall the total variance for any observation is $\sigma^2_y = \sigma^2_B + \sigma^2_{AB} + \sigma^2$
- Finally, recall for any two random variables W and V the definition of correlation is

$$Corr[W, V] = \frac{\mathrm{COV}[W, V]}{\sqrt{\mathrm{V}[W]\,\mathrm{V}[V]}}$$

Same Location, Same Day

$$
\begin{aligned}
Corr\left[y_{ijk_1}, y_{ijk_2}\right] &= \frac{\mathrm{COV}\left[y_{ijk_1}, y_{ijk_2}\right]}{\sqrt{\mathrm{V}\left[y_{ijk_1}\right]\mathrm{V}\left[y_{ijk_2}\right]}} \\
&= \frac{\mathrm{COV}\left[\beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk_1}, \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk_1}\right]}{\sigma_B^2 + \sigma_{AB}^2 + \sigma^2} \\
&= \frac{\mathrm{COV}\left[\beta_j, \beta_j\right] + \mathrm{COV}\left[(\alpha\beta)_{ij}, (\alpha\beta)_{ij}\right]}{\sigma_B^2 + \sigma_{AB}^2 + \sigma^2} \\
&= \frac{\sigma_B^2 + \sigma_{AB}^2}{\sigma_B^2 + \sigma_{AB}^2 + \sigma^2}
\end{aligned}
$$

Not zero! These observations are correlated!

More Correlations!

- What about different locations on the same day?

  $$Corr\left[y_{i_1jk}, y_{i_2jl}\right] = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_{AB}^2 + \sigma^2}$$

  Also non-zero. These observations are correlated too!

- What about observations from different days?

  $Corr = 0$

  These are zero!

- **Conclusion:** Observations taken from within the same level of a random effect are correlated. Across random effects they are uncorrelated.

Estimating and Using Correlations

- For the Chicken Processing Plant Example this yields:
  - **Same Location, Same Day:** $\frac{0.016+0.021}{0.016+0.021+0.55} = \frac{0.037}{.587} = 0.063$
  - **Different Location, Same Day:** $\frac{0.016}{0.016+0.021+0.55} = \frac{0.016}{.587} = 0.027$
- Observations not highly correlated, but certainly not independent - will affect estimation
- What would the correlation (or covariance) matrix of $\mathbf{y}$ look like?

Estimating the Mean Responses

- What's the effect of a non-zero covariance (or correlation)?
  Consider a pairwise contrast for Location 4 vs. Location 3
- $V\left[\bar{y}_{4\cdot\cdot} - \bar{y}_{3\cdot\cdot}\right] \neq \sigma^2\left(\frac{1}{nb} + \frac{1}{nb}\right)$
- How would we estimate a variance (or standard error) here?
  Skipping all the math ... $V\left[\bar{y}_{4\cdot\cdot} - \bar{y}_{3\cdot\cdot}\right] = \frac{2}{nb}\left(\sigma^2 + n\sigma^2_{AB}\right)$
  Easy to estimate! $\widehat{SE}\left[\bar{y}_{4\cdot\cdot} - \bar{y}_{3\cdot\cdot}\right] = \sqrt{\frac{2}{nb}(MSAB)}$
  Exactly a MS from our table - no need for Satterthwaite's formula
- What about individual levels like $\bar{y}_{i\cdot\cdot}$ rather than contrasts?
  Again skipping the math ... $\frac{1}{nb}\left(\sigma^2 + n\sigma^2_B + n\sigma^2_{AB}\right)$
  Algebra yields $\widehat{SE}\left[\bar{y}_{i\cdot\cdot}\right] = \sqrt{\frac{1}{nab}((a-1)MSAB + MSB)}$
  Not exactly a MS - need to use Satterthwaite's formula for DF
  (7.33 for this example...)

Summary: Chicken Processing Plant Example

- Day-to-day variability is not an issue
- Mean log-count is different at earlier locations (1 & 2) than later locations (3 & 4)
- Simple concepts - e.g. confidence interval for a group mean - have become much more complicated
- Understanding the basic process allows us to fit correct model and understand output
- E.g. why DF for intercept and other betas aren't the same?

  E.g., why DF for Location effect and Location Contrasts aren't both affected by DDFM!?