32 Lecture 32: April 14

Last time

- Sample size computations for one-way ANOVA
- Lack of fit test
- One-way random effect model (JF Chapter 23 + Dr. Osborne's notes)

Today

- hypothesis test and confidence intervals for one-way random-effects model
- review of one-way random effects ANOVA model
- nested design

Additional reference

Course notes by Dr. Jason Osborne.
Lecture notes from Lukas Meier on ANOVA using R

Other parameters of interest in random effects models

Coefficient of variation (CV):

$$CV(Y_{ij}) = \frac{\sqrt{Var(Y_{ij})}}{|E(Y_{ij})|} = \frac{\sqrt{\sigma_T^2 + \sigma^2}}{|\mu|}$$

Intraclass correlation coefficient:

$$\rho_I = \frac{Cov(Y_{ij}, Y_{ik})}{\sqrt{Var(Y_{ij})\, Var(Y_{ik})}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2}$$

- Interpretation: the correlation between two responses receiving the same level of the random factor.
- Bigger values of ρ_I correspond to (bigger/smaller?) random treatment effects.

For sires,

$$\widehat{CV} = \frac{\sqrt{117 + 464}}{82.6} = 0.29$$

$$\hat{\rho}_I = \frac{117}{117 + 464} = 0.20$$

Interpretations:

- The estimated standard deviation of a birthweight, 24.1 is 29% of the estimated mean birthweight, 82.6.
- The estimated correlation between any two calves with the same sire for a male parent, or the estimated *intrasire* correlation coefficient, is 0.20.

Testing a variance component - $H_0 : \sigma_T^2 = 0$

Recall that $\sigma_T^2 = Var(T_i)$, the variance among the population of treatment effects.

$$F = \frac{MS[T]}{MS[E]}$$

reject H_0 at level α if $F > F(\alpha, t-1, N-t)$.

For the sires data,

$$F = \frac{1398}{464} = 3.01 > 2.64 = F(0.05, 4, 35)$$

so H_0 is rejected at $\alpha = 0.05$. (The p-value is 0.0309)

Interval estimation of some model parameters

A 95% confidence interval for μ derived by considering $SE(\bar{Y}_{++})$:

$$\begin{aligned}
\bar{Y}_{++} &= \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n} Y_{ij} \\
&= \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n}(\mu + T_i + \epsilon_{ij}) \\
&= \mu + \bar{T}_+ + \bar{\epsilon}_{++}
\end{aligned}$$

where $\bar{T}_+ = (T_1 + \cdots + T_t)/t$ and $\bar{\epsilon}_{++} = (\sum\sum \epsilon_{ij})/N$, so that

$$\begin{aligned}
Var(\bar{Y}_{++}) &= Var(\bar{T} + \bar{E}_{++}) \\
&= \frac{\sigma_T^2}{t} + \frac{\sigma^2}{nt} \\
&= \frac{1}{nt}(n\sigma_T^2 + \sigma^2) \\
&= \frac{1}{nt}E(MS[T]).
\end{aligned}$$

If the data are normally distributed, then

$$\frac{\bar{Y}_{++} - \mu}{\sqrt{\frac{MS[T]}{nt}}} \sim t_{t-1}$$

and a 95% confidence interval for μ is given by

$$\bar{Y}_{++} \pm t(0.025, t-1)\sqrt{\frac{MS[T]}{nt}}$$

For the sires data: $\bar{y}_{++} = 82.6$, $MS[T] = 1398$, $nt = 40$. Critical value $t(0.025, 4) = 2.78$ yields the interval

$$82.6 \pm 2.78(5.91) or (66.1, 99.0).$$

Confidence interval for ρ_I:
A 95% confidence interval for ρ_I can be obtained from the expression

$$\frac{F_{obs} - F_{\alpha/2}}{F_{obs} + (n-1)F_{\alpha/2}} < \rho_I < \frac{F_{obs} - F_{1-\alpha/2}}{F_{obs} + (n-1)F_{1-\alpha/2}}$$

where $F_{\alpha/2} = F(\alpha/2, t-1, N-t)$ and F_{obs} is the observed F-ratio for treatment effect from the ANOVA table.

For the sires data, $F_{obs} = 3.01$ and $F_{0.025} = 3.179$, $F_{0.975} = 0.119$. The formula gives $(-0.01, -0.75)$.

These formulas arrived at via some distributional results:

- $(t-1)\frac{MS[T]}{\sigma^2 + n\sigma_T^2} \sim \chi^2_{t-1}$
- $(N-t)\frac{MS[E]}{\sigma^2} \sim \chi^2_{N-t}$
- $MS[T]$ and $MS[E]$ are independent
- Ratio of independent χ^2 random variables divided by *df* has an F distribution
- $\left(\frac{MS[T]}{\sigma^2 + n\sigma_T^2}\right) / \left(\frac{MS[E]}{\sigma^2}\right) \sim F_{t-1,N-t}$
  (which explains the F test for $H_0 : \sigma_T^2 = 0$)
- Rearranging the probability statement below

$$1 - \alpha = \Pr\left(F(1-\frac{\alpha}{2}, t-1, N-t) < \frac{\frac{MS[T]}{\sigma^2 + n\sigma_T^2}}{\frac{MS[E]}{\sigma^2}} < F(\frac{\alpha}{2}, t-1, N-t)\right)$$

Confidence interval for variance components:

The estimated residual variance component for the sire data was $\hat{\sigma}^2 = MS[E] = 464\ lbs^2$.

A 95% confidence interval for this variance component is given by

$$\left(\frac{(40-5)464}{53.2} < \sigma^2 < \frac{(40-5)464}{20.6}\right)$$

or $(305.2, 789.5)\ lbs^2$

This can be derived using the distributional result

$$(N-t)\frac{MS[E]}{\sigma^2} \sim \chi^2_{N-t}$$

setting up the probability statement

$$1 - \alpha = \Pr\left(\chi^2(1-\frac{\alpha}{2}, N-t) < (N-t)\frac{MS[E]}{\sigma^2} < \chi^2(\frac{\alpha}{2}, N-t)\right)$$

Rearranging to get σ^2 in the middle yields the $100(1-\alpha)\%$ confidence interval for σ^2:

$$\left(\frac{(N-t)MS[E]}{\chi^2_{\alpha/2}}, \frac{(N-t)MS[E]}{\chi^2_{1-\alpha/2}} \right).$$

Question: what are the mean and variance of χ^2_{35} distribution? *Answer:*

Confidence interval for σ_T^2:

The estimated variance component for the random sire effect was $\hat{\sigma}_T^2 = 117$.

Q: How can we get a 95% confidence interval for σ_T^2?
A: In a similar fashion, but the confidence level based on Satterthwaite's approximation to the degrees of freedom of the linear combination of MS terms:

$$\left(\frac{\widehat{df}\hat{\sigma}_T^2}{\chi^2_{\alpha/2,\widehat{df}}}, \frac{\widehat{df}\hat{\sigma}_T^2}{\chi^2_{1-\alpha/2,\widehat{df}}} \right)$$

where

$$\widehat{df} = \frac{(n\hat{\sigma}_T^2)^2}{\frac{MS[T]^2}{t-1} + \frac{MS[E]^2}{N-t}}$$

For the sire data,

$$\widehat{df} = \frac{(8 \times 117)^2}{\frac{1398^2}{4} + \frac{464^2}{35}} = 1.76$$

and

$$\chi^2_{0.975,1.79} = 0.029,\ \chi^2_{0.025,1.76} = 6.87$$

yielding the 95% confidence interval

$$\left(\frac{1.76(117)}{6.87}, \frac{1.76(117)}{0.29} \right)$$

or

$$(30, 7051)$$

Review of one-way random effects ANOVA

The one-way random effects model

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{T_i}_{\text{random}} + \underbrace{\epsilon_{ij}}_{\text{random}} \qquad \text{for } i = 1, 2, \ldots, t \text{ and } j = 1, \ldots, n$$

with

- $T_1, T_2, \ldots, T_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2)$
- $\epsilon_{11}, \ldots, \epsilon_{tn} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
- $T_1, T_2, \ldots, T_t$ independent of $\epsilon_{11}, , \ldots, \epsilon_{tn}$

Remarks:

- $T_1, T_2, \ldots$ randomly drawn from population of treatment effects.
- Only three parameters: μ, σ^2, and σ_T^2
- Several functions of these parameters of interest
  - Coefficient of variation: $CV(Y) = \frac{\sqrt{\sigma^2+\sigma_T^2}}{\mu}$
  - Intraclass correlation coefficient: $\rho_I = Corr(Y_{ij}, Y_{ik}) = \frac{\sigma_T^2}{\sigma^2+\sigma_T^2}$
- Two observations from same treatment group are **not** independent

Exercise: match up the formulas for confidence intervals below with their targets, ρ_I, σ^2, σ_T^2, μ:

$$\bar{Y}_{++} \pm t(0.025, t-1)\sqrt{\frac{MS[T]}{nt}}$$

$$\left(\frac{F_{obs} - F_{\alpha/2}}{F_{obs} + (n-1)F_{\alpha/2}}, \frac{F_{obs} - F_{1-\alpha/2}}{F_{obs} + (n-1)F_{1-\alpha/2}}\right)$$

$$\left(\frac{(N-t)MS[E]}{\chi^2_{\alpha/2}}, \frac{(N-t)MS[E]}{\chi^2_{1-\alpha/2}}\right)$$

$$\left(\frac{\widehat{df}\hat{\sigma}_T^2}{\chi^2_{\alpha/2,\widehat{df}}}, \frac{\widehat{df}\hat{\sigma}_T^2}{\chi^2_{1-\alpha/2,\widehat{df}}}\right)$$

Modelling factorial effects: fixed, or random?

	Random	Fixed
Levels		
- selected from conceptually ∞ population of collection of levels	X	
- finite number of possible levels		X
Another experiment		
- would use same levels		X
- would involve new levels sampled from same population	X	
Goal		
- estimate variance components	X	
- estimate longrun means		X
Inference		
- for these levels used in this experiment		X
- for the population of levels	X	

Nested design

Factor B is nested in factor A if there is a new set of levels of factor B for every different level of factor A.

To illustrate the concept of nested design, we consider the "Pastes" data set in "lme4" package in R. The strength of a chemical paste product was measured for a total of 60 samples coming from 10 randomly selected delivery batches each containing 3 randomly selected casks. Hence, two samples were taken from each cask. We want to check what part of the variability of strength is due to batch and cask.

Let Y_{ijk} be the strength of the kth sample of cask j in batch i. We can use the model

$$Y_{ijk} = \mu + A_i + B_{j(i)} + \epsilon_{ijk}$$

where A_i is the random effect of batch and $B_{j(i)}$ is the random effect of cask **within** batch. Note the special notation $B_{j(i)}$ emphasizes that cask is nested in batch.