BIOS 755: Random effects ANOVA

Alexander McLain

ANOVA

- Used to compare the means of some numeric dependent variable for k populations.
- For now, we'll assume we have independent samples of size n_i from each of these populations.
- The k populations are often thought of as being associated with k levels (or treatments) of some factor, which is simply a categorical independent variable.
- This data structure is known as a one-way layout, and the analysis procedure is known as one-way (or one- factor) ANOVA.

Data notation

- Data structure:

POP_1	POP_2	$\dots$	POP_k
y_{11}	y_{21}	$\cdots$	y_{k1}
y_{12}	y_{22}	$\cdots$	y_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{1n_1}	y_{2n_2}	$\cdots$	y_{kn_k}

- y_{ij} = the value of the jth observation in population i,
- $i = 1, 2, \dots, k$, k total populations.
- $j = 1, 2, \dots, n_i$, n_i observations in the ith population.

One-way (One-factor) ANOVA

- We want to test the hypotheses:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \quad vs. \quad H_A : \text{at least one } \mu_i \neq \mu_k$$

- We sample n_i observations at each level, $i = 1, \ldots, k$.
- The model can be written as

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

where $\epsilon_{ij} \sim N(0, \sigma^2)$, $j = 1, \ldots, n_i$.

One-way (One-factor) ANOVA

- The **treatment effects** can be written as $\tau_i = \mu_i - \mu$.
- The statistical model can be written as

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

- Or in vector form

$$\boldsymbol{Y}_i = \mu + \tau_i + \boldsymbol{\epsilon_i}$$

  where $\boldsymbol{\epsilon_i}$ has a diagonal covariance matrix.
- The ANOVA hypotheses can then be written:

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_k = 0 \quad vs. \quad H_A : \text{at least one } \tau_i \neq 0$$

Random Effects in ANOVA

- In the previous ANOVA, the effects are fixed effects since the levels of the factors are the only ones of interest.
- However, if a factor has a very large number of possible levels, we may decide to randomly select a subset of these levels.
- Simplest example: repeated measures, where more than one (identical) measurement is taken on the same individual.
- In this case, the "group/treatment" effect τ_i is best thought of as random because we only sample a subset of the entire population of subjects.

When to Use Random Effects?

- A "group" effect is random if we can think of the levels we observe in that group to be samples from a larger population.
- Thus, we might have multiple sources of variation in the data.
  - **Between-level** variability,σ^2_τ
  - **Within-level** variability, σ^2
- With random effects, our inference is not about just those k levels randomly selected, but it applied to all possible levels.

Some quotes on fixed versus random*

- Effects are fixed if they are interesting in themselves or random if there is interest in the underlying population.
- "When a sample exhausts the population, the corresponding variable is fixed; when the sample is a small (i.e., negligible) part of the population the corresponding variable is random." (Green and Tukey, 1960)
- "If an effect is assumed to be a realized value of a random variable, it is called a random effect." (LaMotte, 1983)
- Fixed effects are estimated using least squares (or, more generally, maximum likelihood) and random effects are estimated with shrinkage (linear unbiased prediction in the terminology of Robinson, 1991). This definition is standard in the multilevel modeling literature (see, for example, Snijders and Bosker, 1999, Section 4.2) and in econometrics.

*Taken from Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, 33(1), 1–53. Not edited, cause some of these stink.

Statistical model for random effects ANOVA

- The formal statistical model we are considering can be written as

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

where

  - Y_{ij} = measurement j for "treatment" i.
  - μ overall mean.
  - $\tau_i \sim^{iid} N(0, \sigma^2_\tau)$ unobservable random variables
  - $\epsilon_{ij} \sim^{iid} N(0, \sigma^2)$ and independent of the τ_i's.

- Also called "Components of Variance Model" or "Model II ANOVA", or (a simple) "Hierarchical (or Multilevel) Model."

Intraclass correlation

- The total variability in the data is $\hat{\sigma}^2 + \hat{\sigma}^2_\tau$.
- The proportion of variability that is attributed to the between-treatment variability is

$$\text{ICC} = \frac{\sigma^2_\tau}{\sigma^2 + \sigma^2_\tau}$$

  which is referred to as the intraclass correlation (ICC).
- This is estimated by

$$\widehat{\text{ICC}} = \frac{\hat{\sigma}^2_\tau}{\hat{\sigma}^2 + \hat{\sigma}^2_\tau}$$

Random Effects the Concepts

- In public health studies, we randomly select **k** people from a population to be in a study.
- In this circumstance we have two sources of variability
  - person to person variability (between treatment=σ^2_τ).
  - within person variability (within treatment=σ^2).
- Previously, we've broken the within-person variability down further into measurement error and real individual fluctuations in the mean not picked up by our model. Right now, we'll group those together.

Random Effects the Concepts

- For example, we might be looking at blood pressure from 5 individuals at 10 time points.
  - Variability in mean blood pressure between the 5 individuals = between treatment variability = σ_τ^2.
  - Variability in blood pressure measurements at the 10 time points = within treatment variability = σ^2.

Random Effects ANOVA example

Go to example.

Hypothesis Testing with Random Effects

- With fixed effects we are interested in testing

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_k$$

but there we 'know' that these k levels are all the levels there are.

- With random effects we are not just interested the k levels we observed (i.e., the 5 subjects).
- We are interested in testing if there is a significant difference between *all possible levels* (i.e., in all possible subjects).
- As a result, we are interested in testing

$$H_0 : \sigma^2_\tau = 0$$

i.e., variability in the *random effect* doesn't impact the outcome (use `covtest` in SAS).

RE ANOVA

- The RE ANOVA model (which we've been using) does not have correlation for e_i. The model is

$$Y_{ij} = \mu + \tau_i + e_{ij}$$

where

  - $\tau_i \sim^{iid} N(0, \sigma^2_\tau)$ unobservable random variables
  - $e_{ij} \sim^{iid} N(0, \sigma^2)$ and independent of the τ_i's.

- **What gives?**

RE ANOVA as a GLM

- It turns out that

$$Y_{ij} = \mu + \tau_i + e_{ij}$$

where $\tau_i \sim^{iid} N(0, \sigma^2_\tau)$ and $e_{ij} \sim^{iid} N(0, \sigma^2)$.

- Is the same as

$$Y_{ij} = \mu + e_{ij}$$

when

$$Corr(e_{ij}, e_{ik}) = \frac{\sigma^2_\tau}{\sigma^2 + \sigma^2_\tau}.$$

RE ANOVA as a GLM

- And the $Corr(e_{ij}, e_{ik}) = \sigma^2_\tau/(\sigma^2 + \sigma^2_\tau)$ when

$$Cov(\boldsymbol{e}_i) = \Sigma = \begin{pmatrix} \sigma^2 + \sigma^2_\tau & \sigma^2_\tau & \dots & \sigma^2_\tau \\ \sigma^2_\tau & \sigma^2 + \sigma^2_\tau & \dots & \sigma^2_\tau \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_\tau & \sigma^2_\tau & \dots & \sigma^2 + \sigma^2_\tau \end{pmatrix}$$

which is call a **Compound Symmetric** covariance model.