Stat 414 – Day 6
Random Intercepts and Induced Correlation (Ch. 4)

Last Time:

- Traditional ANOVA vs. anova model comparison
- Blocking variables
  - If not included in the analysis, can lead to misleading standard errors
  - $Y_{ij} = \beta_0 + \beta_1 x_{ij} + blocking\ effect_j + \epsilon_{ij}$
    - $E(Y_{ij}|person_j) = \beta_0 + \beta_1 x_{ij} + blocking\ effect_j; V(Y_{ij}|person_j) = \sigma^2$
  - Special case: treatment variable: $Y_{ij} = \mu + \alpha_i + p_j + \epsilon_{ij}$
- Random effects
  - $Y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + \epsilon_{ij}$
    - $E(Y_{ij}) = \beta_0 + \beta_1 x_{ij}\ ; V(Y_{ij}) = \tau^2 + \sigma^2$

Example 1: Data were collected from nine beaches along the Dutch coast. Five readings were taken for each beach, measuring the species richness (number of different species). Let's consider the "null model" (the "random intercepts" or "variance components" model).

$Y_{ij} = \beta_0 + u_j + \epsilon_{ij}$ for the i^{th} site on j^{th} beach

(a) We will consider a multilevel model because

(b) Write out as a two-level model

(c) How many Level 1 units and how many Level 2 units are there?

(d) Use R to estimate the variance components (using appropriate notation)

(e) Determine and interpret the intraclass correlation coefficient, ICC (or variance partitioning coefficient, VPC).

What is the corresponding covariance matrix? For beach *j* have 5 sites. The ***covariance matrix*** looks like

	Sites 1	Sites 2	Sites 3…
Sites 1	Var(Y_{1j})	Cov(Y_{1j}, Y_{2j})	Cov(Y_{1j}, Y_{3j})
Sites 2	Cov(Y_{2j}, Y_{1j})	Var(Y_{2j})	Cov(Y_{2j}, Y_{3j})
Sites 3…	Cov(Y_{3j}, Y_{1j})	Cov(Y_{3j}, Y_{2j})	Var(Y_{3j})

(f) What is $Var\left(Y_{1j}\right) = Cov(Y_{1j}, Y_{1j})$? Does this change across beaches?

(g) What is the *covariance* of two different sites on the same beach? Do you expect this value to be positive or negative?

(h) What is the *correlation* of two different sites for the same beach?

(i) What is the covariance between sites for two different beaches?

The intraclass correlation coefficient can be interpreted as

- How much of the variation in the responses is between people
- The correlation of two measurements on the same person

(j) Use a likelihood ratio test to decide whether the variation among beaches is statistically significant.

The large intraclass correlation coefficient tells us that it would not be reasonable to consider the observations within the same beach as independent pieces of information. So do we have 9 observations or 45 observations? In other words, what is the *effective sample size*?

Definition: (Section 3.4)
With the same number of observations in each group, the *design effect* is 1 + (*n* – 1)ICC.
The effective sample size is
(number of groups)(number of observations within groups)/design effect.

(k) What happens when ICC = 0? When ICC = 1?

(l) What is the effective sample size for this study?