Experiments to Study Variances

Suppose we collect 10 soybean samples from randomly selected locations in a field. Each sample is split into two parts, and the protein content of each part is measured to give two protein content measurements for each sample. We are interested in understanding the variability of the protein content measurements among samples and within multiple measurements on each sample. Imagine that the data below are from 2 measurements on each of the 10 selected samples.

Sample	1	2	3	4	5	6	7	8	9	10
Measurement 1	47.9	55.4	53.0	42.2	57.1	40.8	43.9	64.9	52.6	64.0
Measurement 2	48.1	60.0	51.4	46.2	56.3	40.5	44.4	67.2	53.1	59.4

Our model for the measurements is

$$y_{ij} = \mu + a_i + e_{ij} \qquad i = 1, \ldots, t \qquad j = 1, \ldots, r_i$$

- μ is an unknown constant that represents the mean of all measurements.
- The a_i terms are independent $N(0, \sigma_a^2)$ **random effects** that account for variation of the protein contents from sample to sample.
- The e_{ij} terms are independent $N(0, \sigma_e^2)$ random errors that account for variation in multiple measurements for any particular sample.

We will use σ_y^2 to denote variance of all measurements. According to the model, we have

$$\sigma_y^2 = \sigma_a^2 + \sigma_e^2.$$

The variances σ_a^2 and σ_e^2 are known as the **variance components**. We might like to know several things about these variance components.

How variable is our measurement method?

Estimate σ_e^2

Which are more variable, protein contents of samples or measurements of protein contents on a single sample?

Estimate $\dfrac{\sigma_a^2}{\sigma_e^2}$

What proportion of variation in all protein content measurements is due to variation in sample protein content?

Estimate the **intraclass correlation coefficient** $= \rho_I = \dfrac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$.

Note that a measurement method will have $\sigma_e^2 = 0$ and $\rho_I = 1$ if multiple measurements on the same sample are always identical. On the other hand ρ_I will be close to zero if the measurement method gives highly variable readings for each sample.

We can get information about the variance components from the ANOVA table.

Source	DF	Sums of Squares	Mean Squares	Expected Mean Squares
Among	$t-1$	$\sum_{i=1}^{t} r_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$\frac{SSA}{t-1}$	$\sigma_e^2 + r_0\sigma_a^2$
Within	$N-t$	$\sum_{i=1}^{t}\sum_{j=1}^{r_i}(y_{ij} - \bar{y}_{i\cdot})^2$	$\frac{SSW}{N-t}$	σ_e^2
Total	$N-1$	$\sum_{i=1}^{t}\sum_{j=1}^{r_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2$		

$$N = \sum_{i=1}^{t} r_i \qquad\qquad r_0 = \frac{1}{t-1}\left(N - \frac{1}{N}\sum_{i=1}^{t} r_i^2\right)$$

(Note that $N = tr$ and $r_0 = r$ if $r_1 = \cdots = r_t$.)

1. Based on the expected mean squares, what is a reasonable estimate for σ_e^2?

2. Based on the expected mean squares, what is a reasonable estimate for σ_a^2?

3. Based on the expected mean squares, what is a reasonable estimate for $\frac{\sigma_a^2}{\sigma_e^2}$?

4. Based on the expected mean squares, what is a reasonable estimate for ρ_I?

Formula for 95% Confidence Intervals

$$\frac{SSW}{\chi^2_{0.025,(N-t)}} < \sigma_e^2 < \frac{SSW}{\chi^2_{0.975,(N-t)}} \qquad \frac{1}{r_0}\left(\frac{F_0}{F_{0.025,(t-1),(N-t)}} - 1\right) < \frac{\sigma_a^2}{\sigma_e^2} < \frac{1}{r_0}\left(\frac{F_0}{F_{0.975,(t-1),(N-t)}} - 1\right)$$

$$\frac{F_0 - F_{0.025,(t-1),(N-t)}}{F_0 + (r_0-1)F_{0.025,(t-1),(N-t)}} < \rho_I < \frac{F_0 - F_{0.975,(t-1),(N-t)}}{F_0 + (r_0-1)F_{0.975,(t-1),(N-t)}} \qquad F_0 = \frac{MSA}{MSW}$$

In addition to estimating variance components, we might be interested in testing $H_0 : \sigma_a^2 = 0$ against $H_A : \sigma_a^2 > 0$. In our soybean example, the null hypothesis says that the protein content is exactly the same for all soybean samples in the entire field. According to the null hypothesis, the only reason we saw different protein content measurements for the 10 samples we selected is because of variation in the measurement process. The alternative says that not all soybean samples in the field have identical protein contents. This test can be carried out by comparing $F_0 = \frac{MSA}{MSW}$ to the F distribution with $t-1$ and $N-t$ degrees of freedom (just like when testing the equality of means in one-way ANOVA).

5. Given that the sum of squares among soybean samples is 1224.712 and the sum of squares within samples is 33.72, estimate σ_a^2/σ_e^2, estimate ρ_I, find a 95% confidence interval for σ_e^2, and test $H_0 : \sigma_a^2 = 0$ against $H_A : \sigma_a^2 > 0$.