ST 512 - Lab 9 - Introduction to Random Effects

1. Open the `Lab9.sas` file which uses the `Beer` data set. This data set contains measurements on sodium contents in lager for beer from six randomly selected brands of beer produced in the U.S. and Canada. For each beer, eight 12-ounce samples were measured for sodium content.

2. Begin by running the code in Code Block 1 to create a graph of sodium content for each brand.

3. Examine your plot. Which is the bigger source of variance: random effect between brands or random effect due to bottles within brand? Is your answer consistent with what you would've expected?

   The random effect between brands is clearly larger since the points within a brand are relatively tightly clustered, but the points representing each brand are much farther apart. This is exactly what we should have expected - the production company's quality control process should be on point, which would result in small bottle-to-bottle variability.

4. Write the model that treats brand as a random effect. Be able to identify each component (including subscripts) using terms like 'sodium content', 'brand', and 'bottle'. Be sure to include the model assumptions.

   $\mathbf{y_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\epsilon}_{ij}}$

   $\boldsymbol{\tau}_\mathbf{i} \overset{\mathbf{iid}}{\sim} \mathrm{N}\left(\mathbf{0},\ \boldsymbol{\sigma}^2_{\mathbf{Trt}}\right)$

   $\boldsymbol{\epsilon}_\mathbf{ij} \overset{\mathbf{iid}}{\sim} \mathrm{N}\left(0,\ \boldsymbol{\sigma}^2\right)$

   μ **is the true mean sodium content for all brands of beer in US and Canada**

   τ_i **is the effect due to randomly selecting beer brand i**

   ϵ_{ij} **is the effect/residual/error due to randomly selecting bottle j of brand i**

5. Run the code shown in Code Block #2. What output is provided that we didn't see in our previous fixed effects models?

   PROC GLM is now printing the Type III Expected Mean Squares

6. Can you estimate the variance components based on your `PROC GLM` output?

   Yes. By equating mean squares we can find that $\hat{\boldsymbol{\sigma}}^2 = \mathrm{MSE} = 0.716$ and $\hat{\boldsymbol{\sigma}}^2_{\mathbf{Trt}} = \frac{\mathbf{MSTrt - MSE}}{\mathbf{n}} = \frac{\mathbf{170.906 - 0.716}}{\mathbf{8}} = \mathbf{21.27}$

7. Code Block #3 introduces a new procedure `PROC MIXED` that is better equipped to handle random effects. Look at the `MODEL` and `RANDOM` statements used in Code Block #2 vs. #3. What do you think the code in Code Block #3 is doing?

   The MODEL statement in PROC MIXED is only for fixed effects. Random effects go only in the RANDOM statement, unlike PROC GLM.

8. Run the code shown in Code Block #3. How does this output compare to what you saw in Code Block #2?

   The layout is different, and the total line is missing from the ANOVA table, but the output values are identical.

9. Estimate the variance components based on your `PROC MIXED` output in Code Block #3. Do they support you graph-based decision from above about which source of variation was larger?

   Yes - the error variance is approximately 30 times smaller than the brand-related variance.

10. Utilize the expected mean squares table to calculate an estimate for the Expected Mean Square for brand. Is that number anywhere in your output? Does this affect your answer as to whether variance components estimates can be found in the `PROC GLM` output?

    $\boldsymbol{\hat{\sigma}^2 + 8\hat{\sigma}^2_{\text{Trt}} = 170.906}$

11. Run the code shown in Code Block #4. How does this output compare to what you saw in Code Block #3? Do you think that will always happen?

    They are the same. This will not always happen - it depends on when the Methods of Moments and REML methods are equivalent.

12. What hypotheses would be tested to determine if there is any significant brand variability? Test these hypotheses using your output and provide a conclusion.

    $\mathbf{H_0} : \boldsymbol{\sigma^2_{\text{Trt}} = 0}$ **vs.** $\mathbf{H_1} : \boldsymbol{\sigma^2_{\text{Trt}} > 0}$