Stat 414 - Day 14
Random vs. Fixed Effects

Last Time:

- Observations within groups/clusters/subjects are often correlated with each other.
- The intraclass correlation coefficient is one way to measure that correlation. Similar to R^2 it sees how much of the total variation is between groups/clusters/subjects (how reliably can we identify which group an observation is from) and can be intepreted as the correlation between two randomly selected observations from the same group/cluster/subject.
- If you have a larger intraclass correlation coefficient, the effective sample size is smaller.
    - Day 13 typo! $\frac{I \times n}{(n-1) \times ICC + 1}$

Example 1: Caffeine cont. When we looked at just the participants variable, the ICC was 0.7877. We also have $\bar{y}$ = 484 taps/minute(?) and σ_y^2 = 607.45. The OLS fitted model:

Predicted taps = 474 – 17 P1 – 15 P2 – 4 P3 (+35 P4), $\hat{\sigma} = 12.27$
(3.54)

(a) Run the following generalized linear model with "compound symmetry" (equal variances and equal covariances). What is the estimated intercept? Why? Its standard error? How many parameters are estimated? What is $\hat{\sigma}$? Summarize what you learn from the variance-covariance matrix.

```
modelC <- gls(Taps ~ 1 + participant, corr = corCompSymm(form = ~ 1 | participant))
nlraa::var_cov(modelC)
```

(b) How do I convert Cov(W, Z) into Cor(X, Z)?

(c) Is this correlation coefficient significantly different from zero? Try a likelihood ratio test.

```
modelD <- gls(Taps ~ 1, data=fingertapstudy)
nlraa::var_cov(modelD)
anova(modelC, modelD)
```

So then we included participant in the model and found a statistically significant subject-adjusted association between stimulant type and finger tapping rate. But this study is a good example where we aren't really all that interested in the four participants themselves, we were just trying to control for that person-to-person variability, to help us assess the person-adjusted differences among the stimulants. In fact, we might be willing to consider the participants as a random sample?…

(d) Suppose we had a larger study with lots more participants. What would be a downside to including the participant variable in the model?

In a situation like this, one option is to treat person as a *random effect* rather than a *fixed effect*. This means we are going to treat these 4 participants not as 4 levels of a factor, but as a random sample from a population (if I did the study again, I would get 4 different participants). The assumption we are going to make is that the "participant effects" follow a normal distribution, centered at zero, with variance τ^2. Let's call these participant effects, u_j, so we have $u_j \sim N(0, \tau^2)$. Our model equation becomes: $Y_{ij} = \beta_0 + u_j + \epsilon_{ij}$ where $u_j \sim N(0, \tau^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$. (for the i^{th} observation on the j^{th} subject)

Big deal, I changed β's to u's, but that is one way of saying we aren't considering the participant effects as parameters anymore. Instead, we replace them with one parameter, τ^2, which represents the participant-to-participant variation in the population of (potential) participants. This "small" change will have a large impact on the properties of the model.

To fit this model, today we will use the "lme" command from the nlme package which you already have because it contains gls.

```
model4 = lme(fixed = Taps ~ 1 , random = ~1 | participant, data = fingertapstudy, method="REML")
#(1|subject) is how we tell R to treat the participants as random effects
```

(e) How many parameters are estimated in this model? How does the estimated intercept change? Standard error? What are the estimated variance components?

We can view the estimated variance-covariance matrix for individual subjects.

```
getVarCov(model4, subject = "1", type = "marginal")[[1]]
```

And we can make R do the conversion to correlations

```
cov2cor(getVarCov(model4, subject = "1", type = "marginal")[[1]])
```

So we have partitioned the total random variability into a variance component for the individual observations within each person (assumed to be the same across the participants) and a variance component for the participants.

(f) Find the estimated "total variation" by summing* $\hat{\tau}^2 + \hat{\sigma}^2$.

(g) How much of this variation is due to the different participants?

Note: Some packages/functions report the estimated variances, some the estimated standard deviations, some both.