Simulation and parametric bootstrap

Recap: Data and Goal

We have date on 197 early-stage *Bugula neritina*, with information on

- `Run`: which repetition of the experiment the individual was recorded in
- `Mass`: Mass of the individual (in micrograms)
- `Metabolic`: Recorded metabolic rate (rate of energy consumption) of the individual (in mJ per hour)

> **Goal for this class:** Is there systematic variation between different runs (i.e., is there any correlation due to Run)?

Plan (so far)

$$Metabolic_{ij} = \beta_0 + \varepsilon_{ij} \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2_{\varepsilon})$$

> How unusual are the observed differences between runs, if there is really no systematic differences between runs (i.e., no random effects)?

- Pretend that the intercept-only model is correct
  - $Metabolic^* = 0.175 + \varepsilon^* \qquad \varepsilon^* \sim N(0, 0.0043)$
- Create a new dataset from the intercept-only model

```
new_metabolic <- 0.175 +
  rnorm(n=197, mean=0, sd=sqrt(0.0043))
```

- Compare our new dataset to the observed dataset

Our simulated data

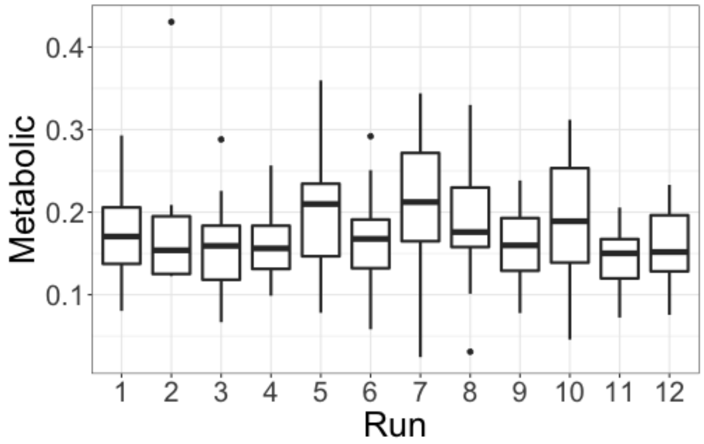
$$Metabolic^*_{ij} = 0.175 + \varepsilon^*_{ij} \qquad \varepsilon^*_{ij} \overset{iid}{\sim} N(0,\ 0.0043)$$

```
new_metabolic <- 0.175 +
  rnorm(n=197, mean=0, sd=sqrt(0.0043))

new_data <- data.frame(Run = bugula_early$Run,
                       Metabolic = new_metabolic)
```

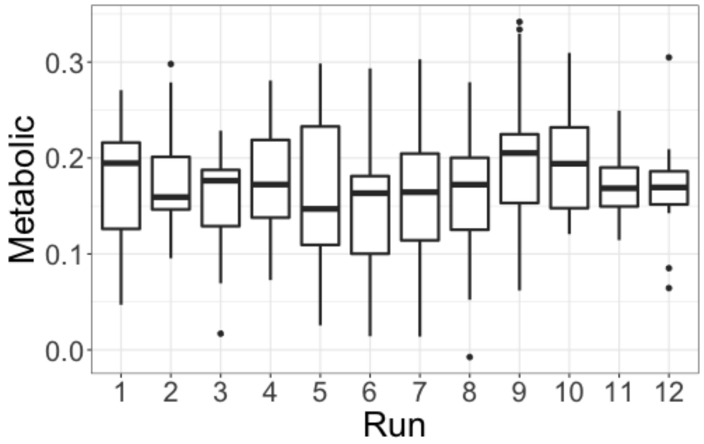
- Create a new metabolic rate for every organism in the data
- Use the same runs from the observed data
- Store the simulated dataset as `new_data`

Compare new dataset to observed dataset

Original (observed) data:



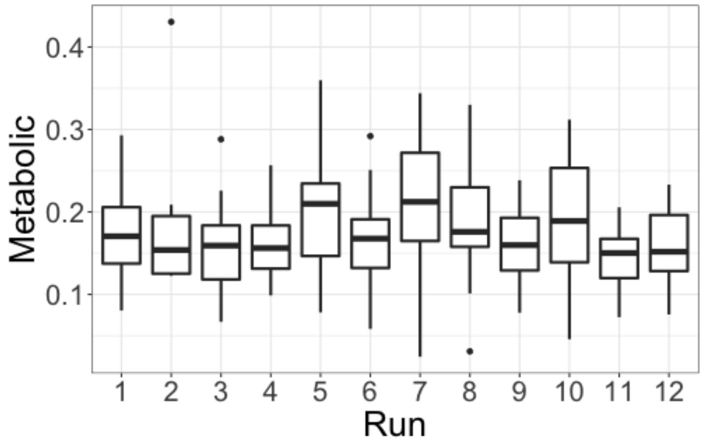
New (simulated) data:



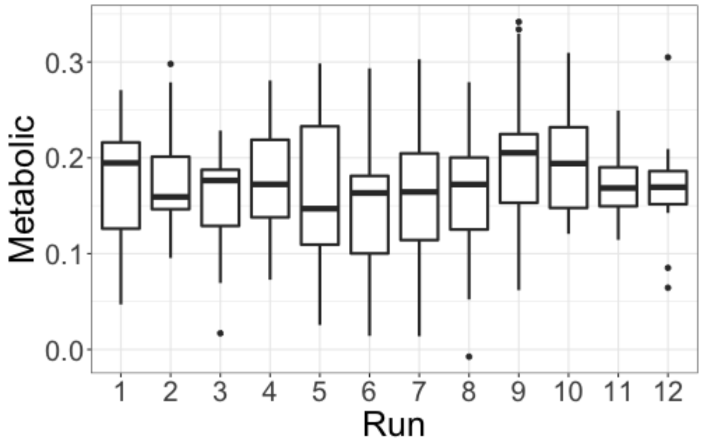
> Do you think there is systematic variation between runs, or do you think the observed differences between runs are due to chance?

Compare new dataset to observed dataset

Original (observed) data:



New (simulated) data:



> How else could I compare the observed data to the simulated data?

Compare new dataset to observed dataset

$$Metabolic_{ij} = \beta_0 + u_i + \varepsilon_{ij} \quad u_i \overset{iid}{\sim} N(0, \sigma_u^2),\ \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

Fitted random intercepts model (observed data):

$$\widehat{\beta}_0 = 0.175,\ \ \widehat{\sigma}_u^2 = 0.00013,\ \ \widehat{\sigma}_\varepsilon^2 = 0.0042,\ \ \hat{\rho}_{group} = 0.03$$

Fitted random intercepts model (simulated data):

$$\widehat{\beta}_0 = 0.169,\ \ \widehat{\sigma}_u^2 = 0.00015,\ \ \widehat{\sigma}_\varepsilon^2 = 0.0049,\ \ \hat{\rho}_{group} = 0.03$$

> Do you think there is systematic variation between runs, or do you think the observed differences between runs are due to chance?

Summary (so far)

> Are there systematic differences between runs (group effects), or are observed differences simply due to chance?

- Fit a model with no random effects
- Simulate data from fitted model
- Compare simulated data to observed data
  - If simulated data looks very different, maybe there are systematic differences between runs
  - If simulated data looks similar to observed data, maybe there aren't systematic differences between runs

Class activity

https://sta214-f22.github.io/class_activities/ca_lecture_31.html

Class activity

```r
m1 <- lmer(na ~ (1|id), data = music)
summary(m1)
```

```
...
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept)  4.95    2.225
##  Residual             22.46    4.739
## Number of obs: 497, groups:  id, 37
...
```

> What is the estimated intra-class correlation?

Class activity

```r
m0 <- lm(na ~ 1, data = music)
summary(m0)
```

```
...
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.2093     0.2349      69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 5.237 on 496 degrees of freedom
...
```

> What are the estimates $\widehat{\beta}_0$ and $\widehat{\sigma}^2_\varepsilon$?

Class activity

$$Anxiety^*_{ij} = \widehat{\beta}_0 + \varepsilon^*_{ij} \qquad \varepsilon^*_{ij} \sim N(0, \widehat{\sigma}^2_\varepsilon)$$

```
new_na <- ... +
  rnorm(n=..., mean=0, sd=...)

new_data <- data.frame(id = music$id,
                       na = new_na)
```

How do I fill in the code to simulate a new dataset from the intercept-only model?

Class activity

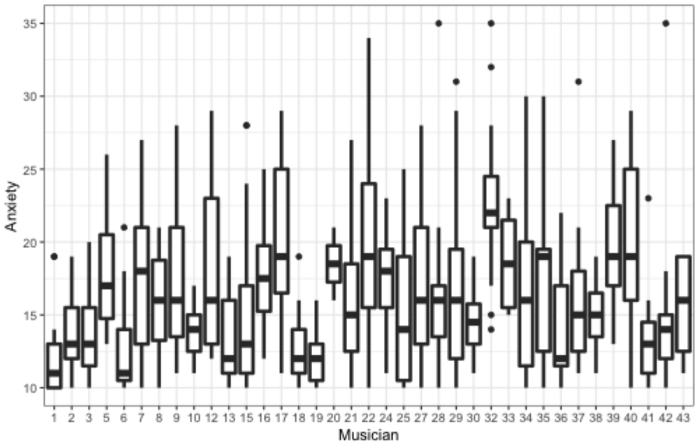
$$Anxiety^*_{ij} = \widehat{\beta}_0 + \varepsilon^*_{ij} \qquad \varepsilon^*_{ij} \sim N(0, \widehat{\sigma}^2_\varepsilon)$$

$\widehat{\beta}_0 = 16.21, \widehat{\sigma}^2_\varepsilon = 5.237^2 = 27.43$

```
new_na <- 16.21 +
  rnorm(n=497, mean=0, sd=5.237)

new_data <- data.frame(id = music$id,
                       na = new_na)
```

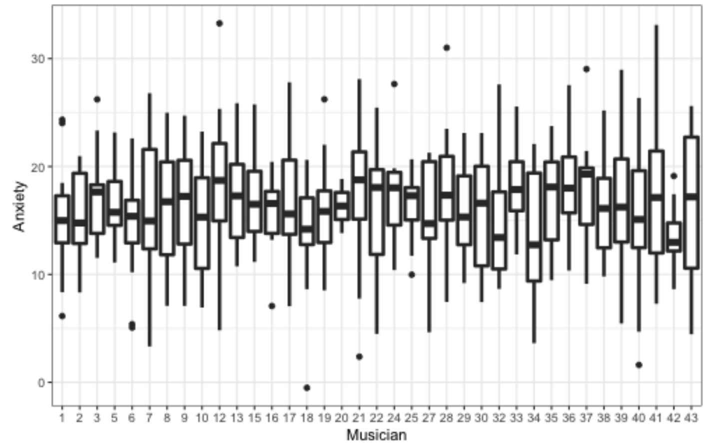
Compare new dataset to observed dataset

Original (observed) data:



$$\hat{\rho}_{group} = 0.18$$

New (simulated) data:



$$\hat{\rho}_{group} = 0$$

Simulating multiple datasets

Plan:

- Simulate a dataset
- Compare it to the observed data (calculate $\hat{\rho}_{group}$)
- Repeat many times (to get a sense of variability)

Simulating multiple datasets

Step 1: Simulate a dataset

```
new_metabolic <- 0.175 +
  rnorm(n=197, mean=0, sd=sqrt(0.0043))

new_data <- data.frame(Run = bugula_early$Run,
                       Metabolic = new_metabolic)
```

> Done!

Simulating multiple datasets

Step 2: Calculate estimated ICC

```
m_sim <- lmer(Metabolic ~ (1|Run),
              data = new_data)

variance_ests <- as.data.frame(summary(m_sim)$varcor)
icc <- variance_ests[1,4]/(variance_ests[1,4] + variance_ests[2,4])

icc
```

```
## [1] 0.009007824
```

- `summary(m_sim)$varcor` extracts variances of the random effect and residuals

Simulating multiple datasets

Step 3: Repeat many times

- First, we need to create a vector to store the results of our simulations

```
nsim <- 200  # do 200 repetitions
iccs <- rep(NA, nsim)  # vector to store the results
```

- `nsim` will be our number of simulated datasets
- `iccs` will store the estimated intra-class correlation for each simulated dataset

> What tool do I use in R to repeat something many times?

Simulating multiple datasets

Step 3: Repeat many times

- Next, we need to iterate with a **for loop**

```
nsim <- 200  # do 200 repetitions
iccs <- rep(NA, nsim)  # vector to store the results

# repeat simulation multiple times
for(sim in 1:nsim){

}
```

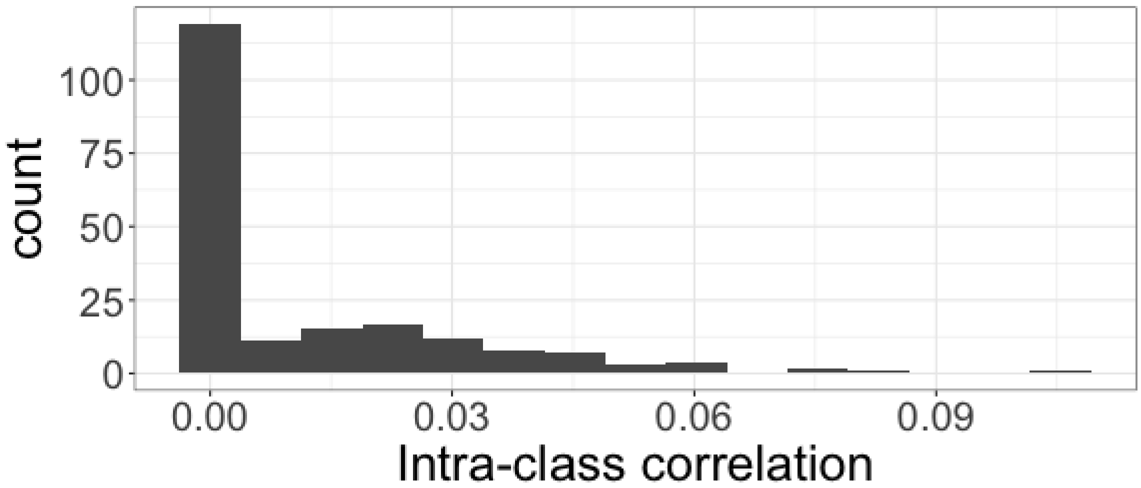
- `for(sim in 1:nsim)` means "repeat what follows `nsim` times"

> What goes inside my for loop?

Simulating multiple datasets

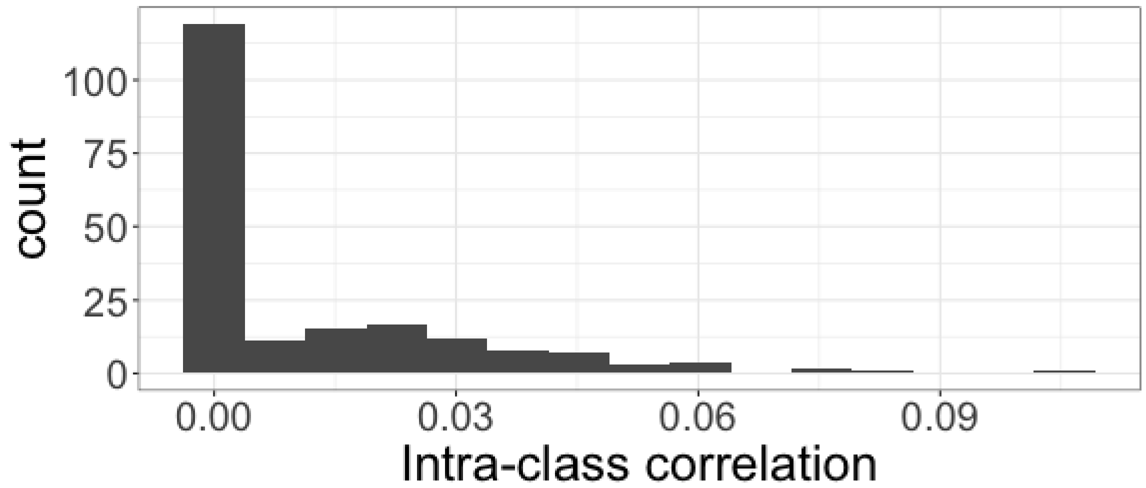
```
nsim <- 200  # do 200 repetitions
iccs <- rep(NA, nsim)  # vector to store the results

# repeat simulation multiple times
for(sim in 1:nsim){
  new_metabolic <- 0.175 +
  rnorm(n=197, mean=0, sd=sqrt(0.0043))

  new_data <- data.frame(Run = bugula_early$Run,
                         Metabolic = new_metabolic)

  m_sim <- lmer(Metabolic ~ (1|Run), data = new_data)

  variance_ests <- as.data.frame(summary(m_sim)$varcor)
  iccs[sim] <- variance_ests[1,4]/(variance_ests[1,4] +
                                   variance_ests[2,4])

}
```

Plotting the results



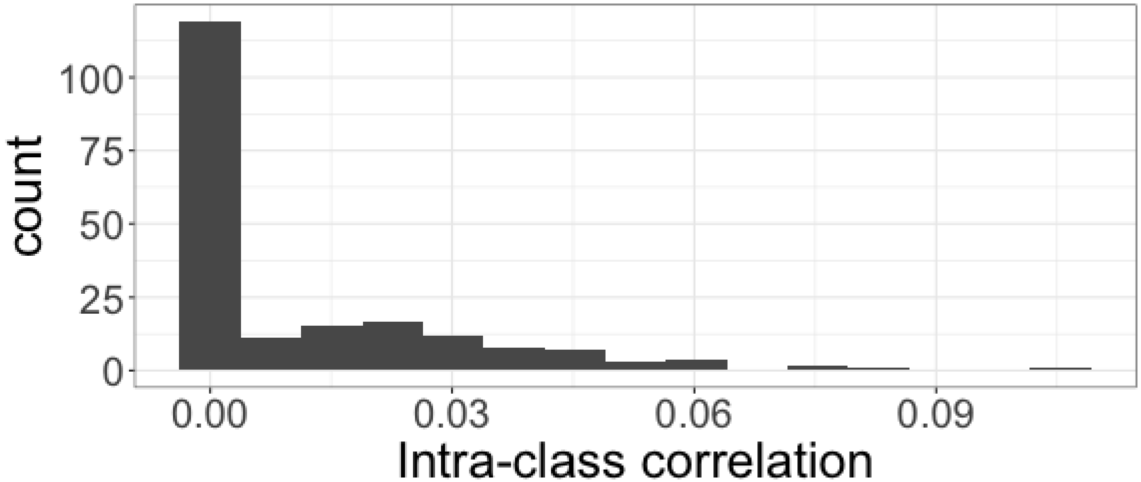
For the observed data, $\hat{\rho}_{group} = 0.03$. Is this unusual, compared to the simulated data?

Plotting the results



How can I summarize how unusual $\hat{\rho}_{group} = 0.03$ is?

Summarizing the results




```
mean(iccs > 0.03)
```

```
## [1] 0.15
```

- The probability of observing $\hat{\rho}_{group}$ as or more extreme than the correlation from the original data, if there is no systematic variation between runs, is about 0.15