Simulation

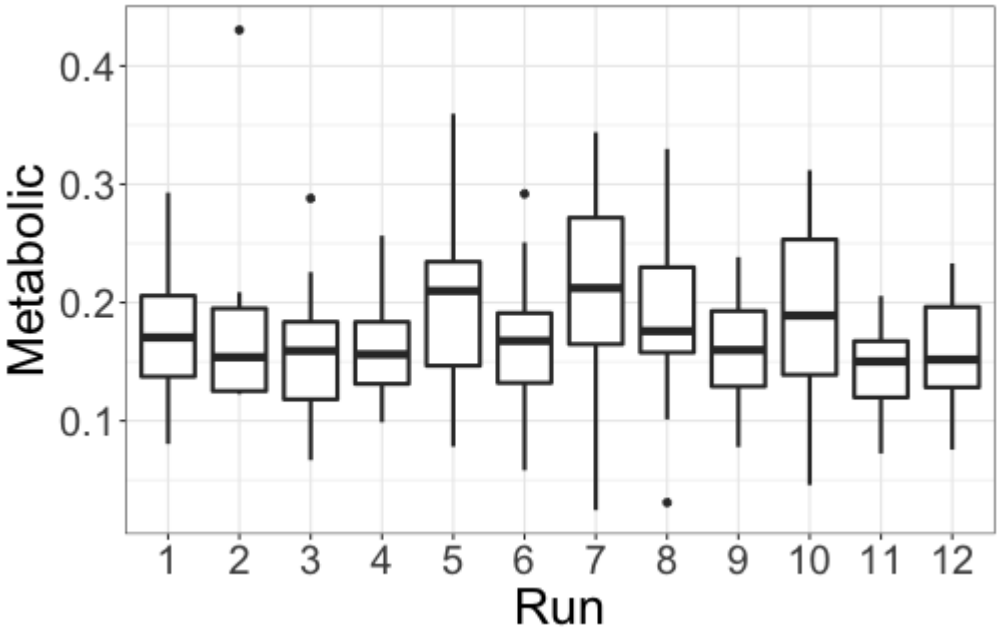
Data and Goal

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)

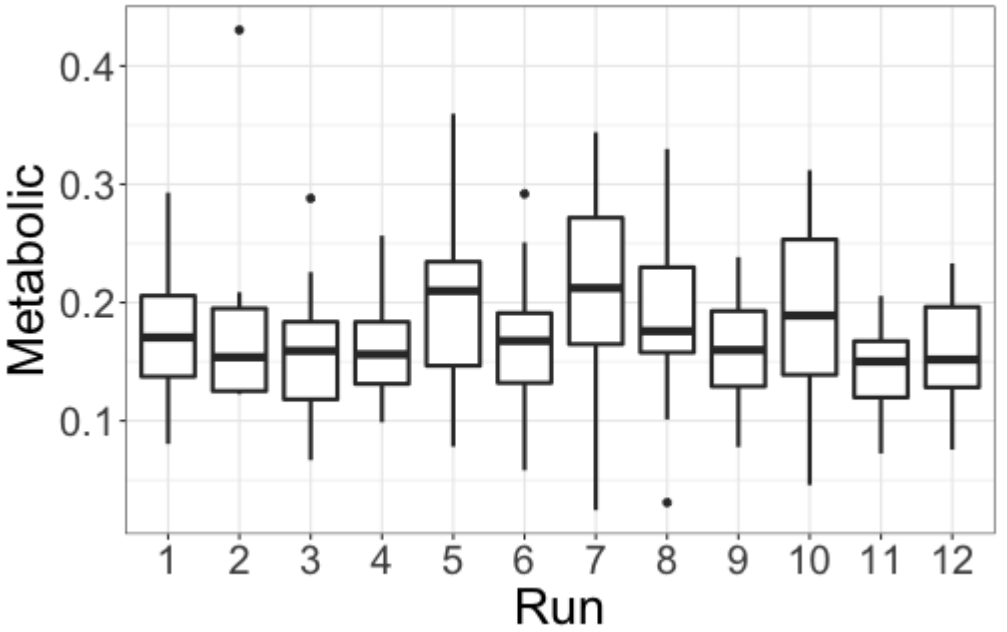
> **Goal for this class:** Is there systematic variation between different runs (i.e., is there any correlation due to Run)?

Visualizing runs



Does it look like there is variation from run to run (i.e., do we need u_i)?

Visualizing runs



It is hard to tell if there are actually differences between runs, or if observed differences are just due to chance.

> How can I estimate variability between runs?

Initial random intercepts model

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

```
m1 <- lmer(Metabolic ~ (1|Run), data = bugula_early)
summary(m1)
```

```
...
##  Groups   Name        Variance Std.Dev.
##  Run      (Intercept) 0.000131 0.01145
##  Residual             0.004181 0.06466
## Number of obs: 197, groups:  Run, 12
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 0.174662   0.005692   30.68
...
```

Initial random intercepts model

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

$$\widehat{\beta}_0 = 0.175, \quad \widehat{\sigma}_u^2 = 0.00013, \quad \widehat{\sigma}_\varepsilon^2 = 0.0042$$

> How would we test whether there is systematic variation between runs?

Plan

Intercept-only model (no random effect)

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

> How do I fit this in R?

Intercept-only model (no random effect)

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

```
m0 <- lm(Metabolic ~ 1, data = bugula_early)
summary(m0)
```

```
...
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.174996   0.004672   37.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.06558 on 196 degrees of freed
...
```

$$\widehat{\beta}_0 = 0.175, \quad \widehat{\sigma}^2_\varepsilon = 0.06558^2 = 0.0043$$

Pretend the intercept-only model is correct

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

> I want to generate a new observation from this relationship. What do I do?

Pretend the intercept-only model is correct

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

$$\widehat{\beta}_0 = 0.175, \quad \widehat{\sigma}^2_\varepsilon = 0.0043$$

- Sample a new noise term: $\varepsilon^* \sim N(0,\ 0.0043)$
- Add the intercept: $0.175 + \varepsilon^*$

So our new observed metabolic rate is

$$Metabolic^* = 0.175 + \varepsilon^*$$

Simulation in R

Sample a new noise term: $\varepsilon^* \sim N(0,\ 0.0043)$

```
rnorm(n=1, mean=0, sd=sqrt(0.0043))
```

```
## [1] -0.06307819
```

Here $\varepsilon^* = -0.063$

Simulation in R

- Sample a new noise term: $\varepsilon^* \sim N(0,\ 0.0043)$
- Add the intercept: $0.175 + \varepsilon^*$

```
0.175 + rnorm(n=1, mean=0, sd=sqrt(0.0043))
```

```
## [1] 0.1119218
```

$Metabolic^* = 0.112$

> In your R console, run the code to generate a new metabolic observation $Metabolic^*$.

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)?

Create a new dataset

Now we sample a new metabolic rate for **every** observation in the data:

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

> How can I modify my R code to sample a new metabolic rate for **every** observation?

```
0.175 + rnorm(n=1, mean=0, sd=sqrt(0.0043))
```

Create a new dataset

Now we sample a new metabolic rate for **every** observation in the 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))
```

Create a new dataset

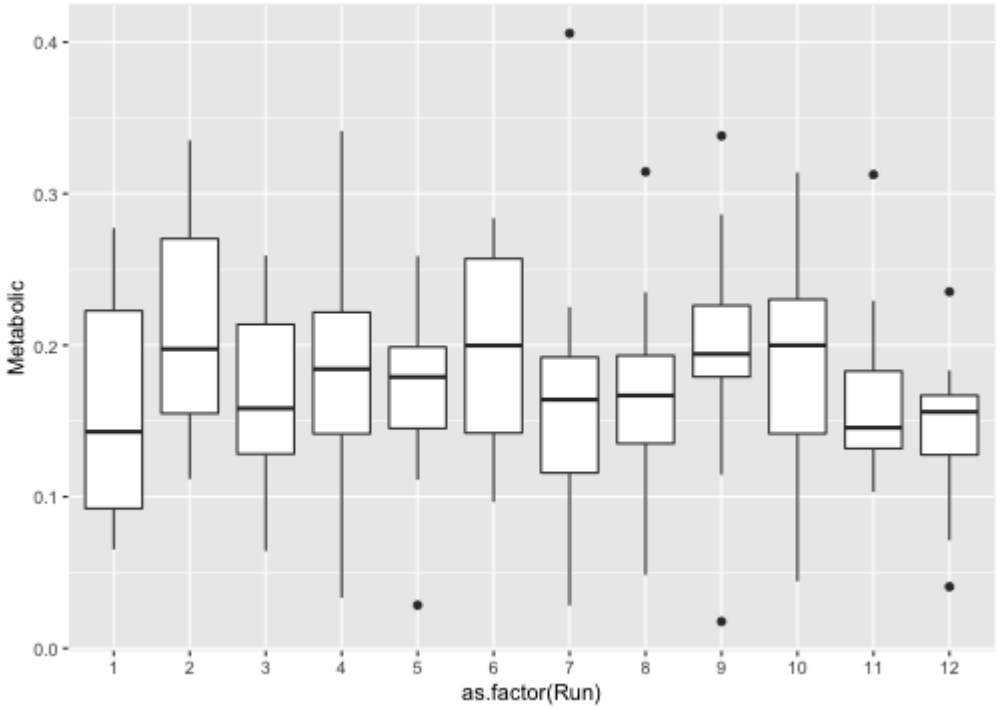
$$Metabolic^*_{ij} = 0.175 + \varepsilon^*_{ij} \quad \varepsilon^*_{ij} \overset{iid}{\sim} N(0,\ 0.0043)$$

```
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`

Create a new dataset

```
new_data %>%
  ggplot(aes(x = as.factor(Run),
             y = Metabolic)) +
  geom_boxplot()
```



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

Class activity

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

Class activity

```
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

```
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

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

```
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

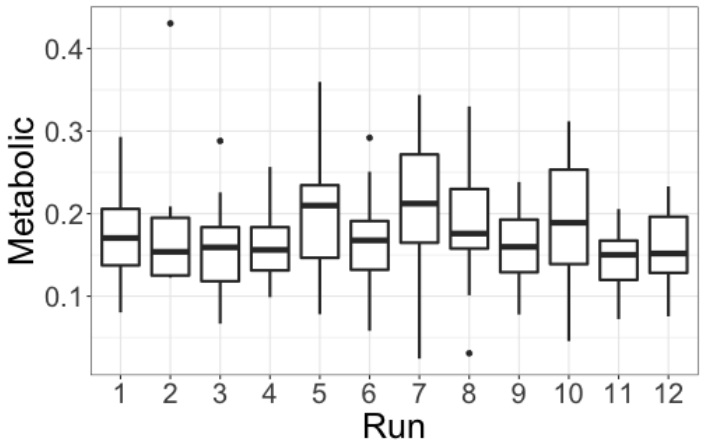
$$Anxiety^*_{ij} = \widehat{\beta}_0 + \varepsilon^*_{ij} \quad \varepsilon^*_{ij} \sim N(0, \widehat{\sigma}^2_\varepsilon)$$

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

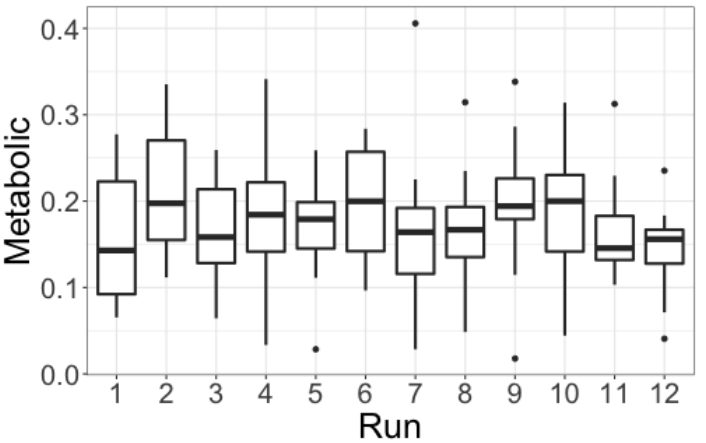
```
new_na <- 16.21 +
  rnorm(n=497, mean=0, sd=5.237)

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

Compare new dataset to observed dataset

Original (observed) data:



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

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

Fitted random intercepts model (observed data):

$$\widehat{\beta}_0 = 0.175, \quad \widehat{\sigma}_u^2 = 0.00013, \quad \widehat{\sigma}_\varepsilon^2 = 0.0042$$

Fitted random intercepts model (simulated data):

$$\widehat{\beta}_0 = 0.169, \quad \widehat{\sigma}_u^2 = 0.00015, \quad \widehat{\sigma}_\varepsilon^2 = 0.0049$$

> 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