Exercise 15

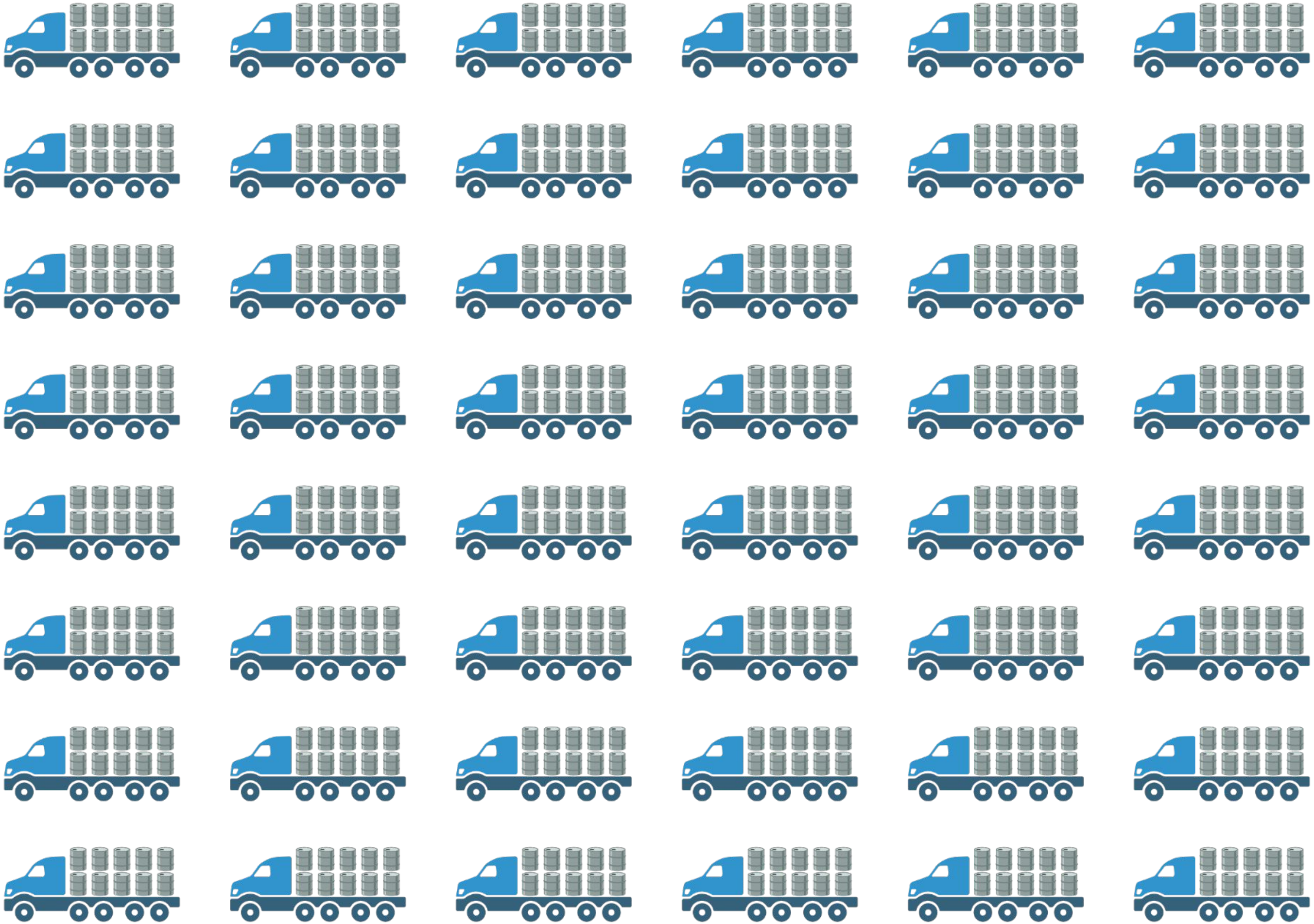
Lea Frei, Laura Fernández and Kristin Watnedal Olsen

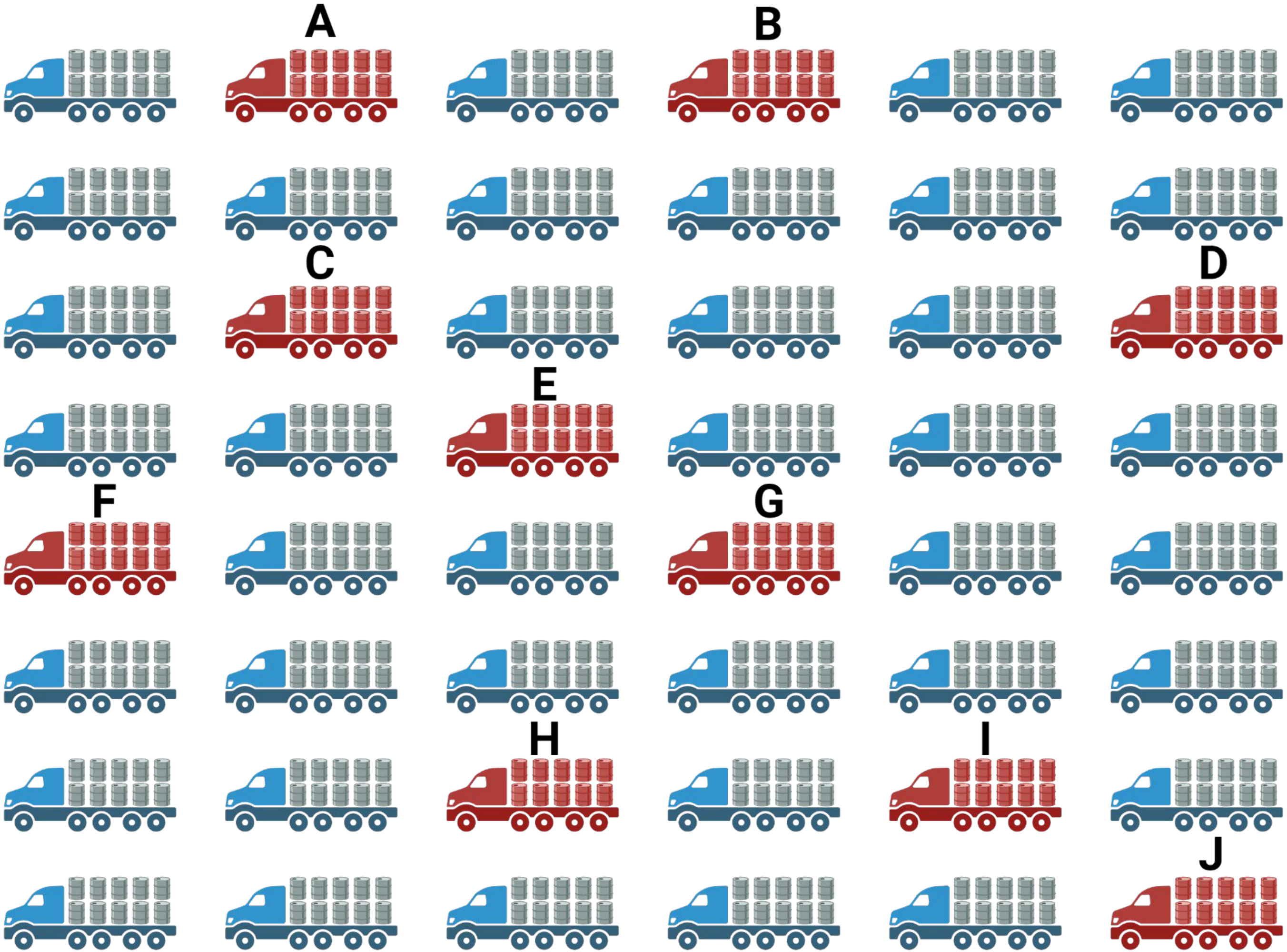
Exercise 15

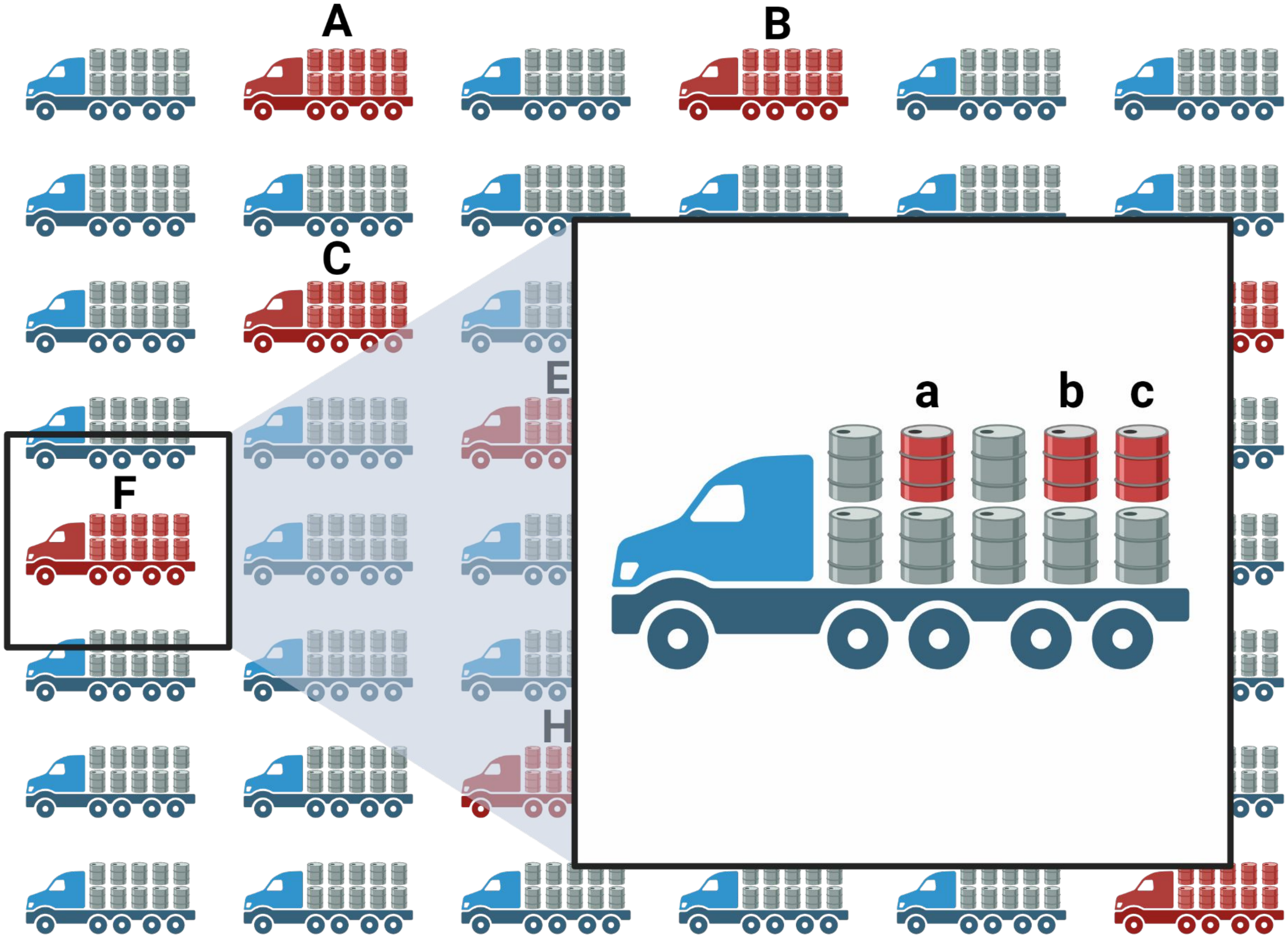
The Pastes data set contains 60 quality measurements (variable strength) of a chemical paste delivered in different batches. From 10 randomly selected delivery batches (variable batch, values ‘A’ to ‘J’) three casks (variable cask, values ‘a’ to ‘c’) were randomly sampled and analyzed twice. This means that we have 30 samples in total (variable sample, values ‘A:a’ to ‘J:c’) and two measurements were carried out on each.

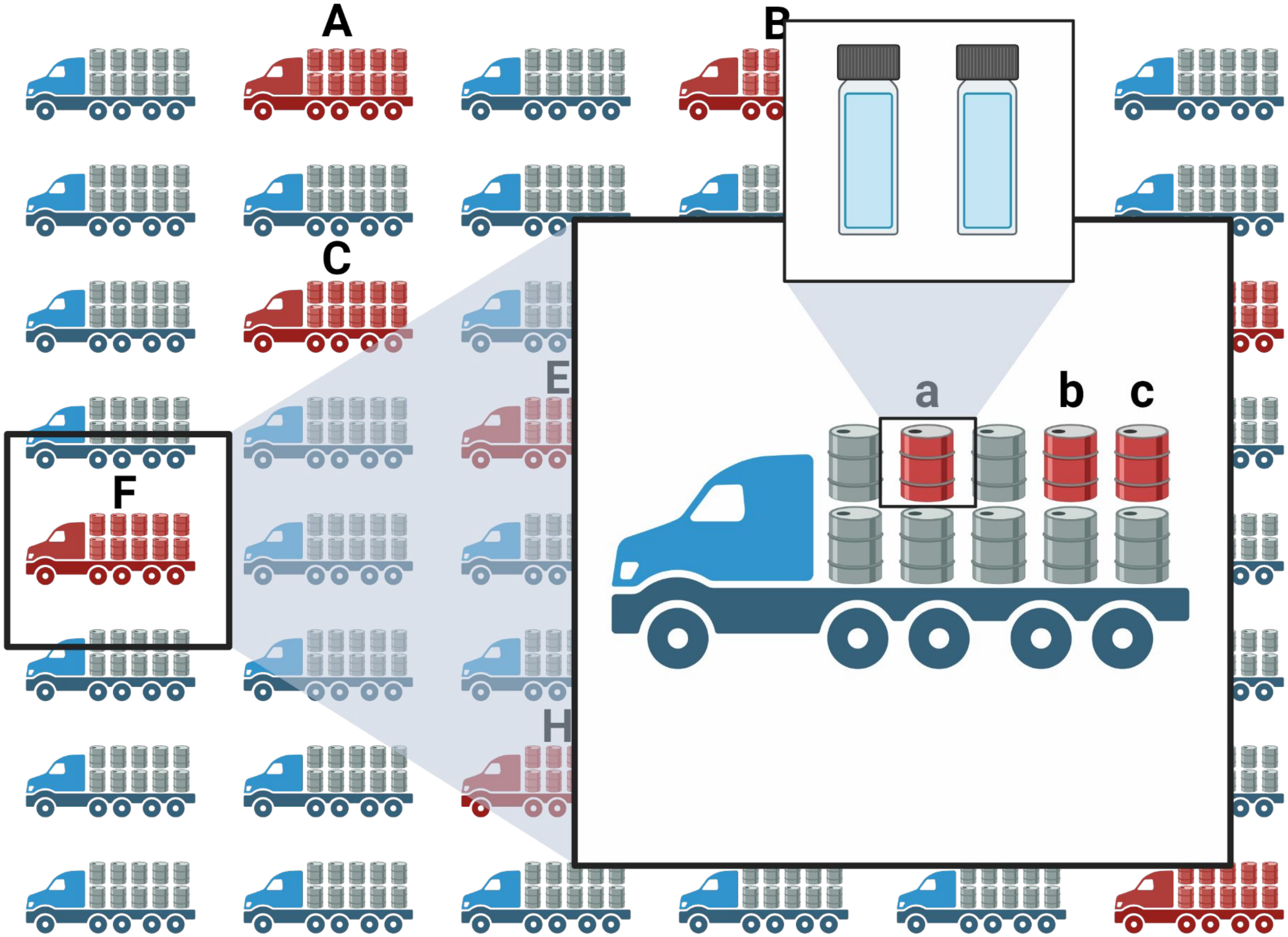
The data set can be loaded as follows:

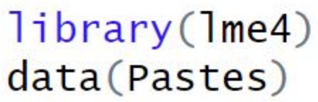
```
> library(lme4)

> data(Pastes)
```



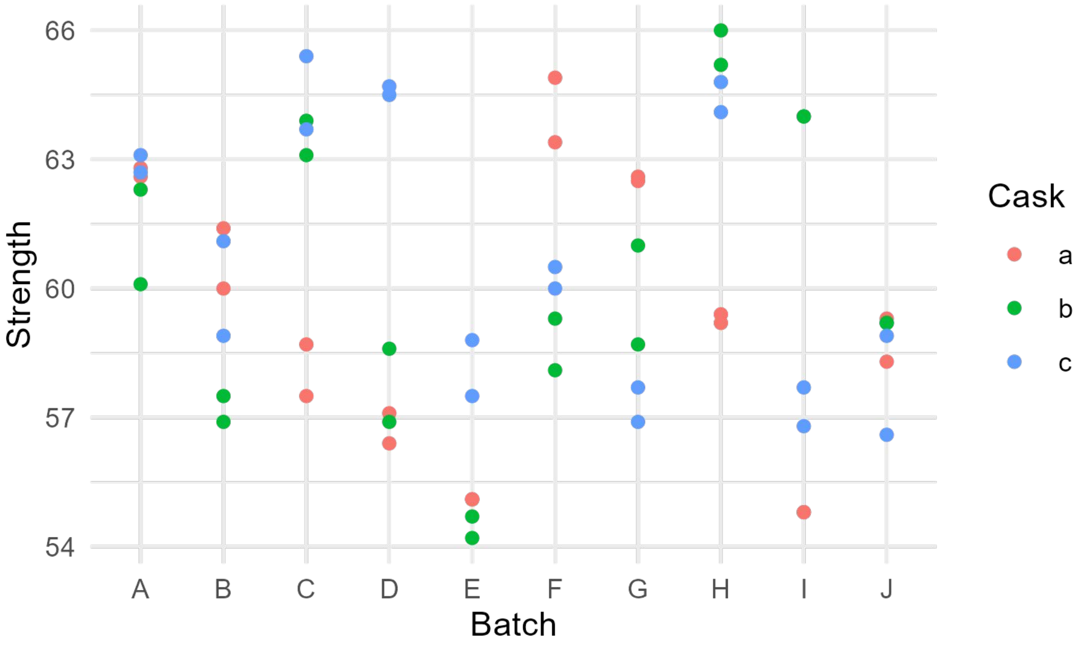

A
B
C
D
E
F
G
H
I
J

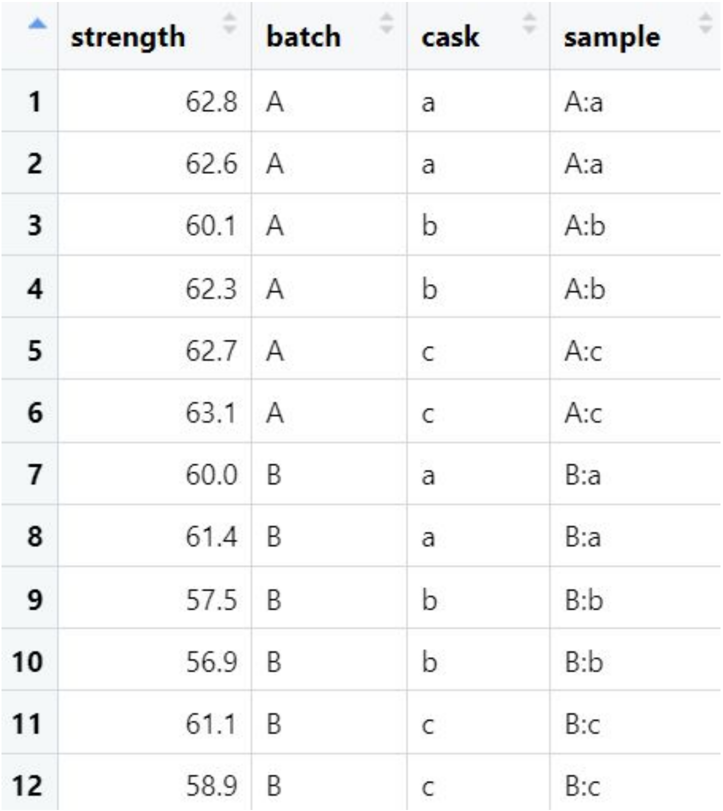

A
B
C
E
F
H
a
b
c


A
B
C
E
F
H
a
b
c

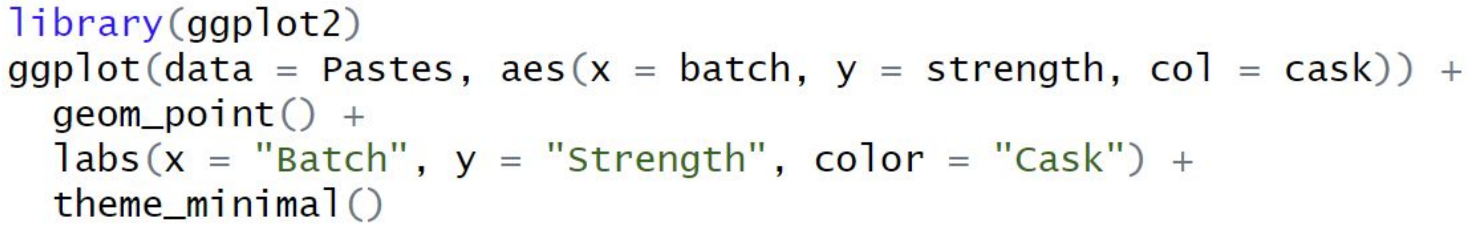


```
library(lme4)
data(Pastes)
```





	strength	batch	cask	sample
1	62.8	A	a	A:a
2	62.6	A	a	A:a
3	60.1	A	b	A:b
4	62.3	A	b	A:b
5	62.7	A	c	A:c
6	63.1	A	c	A:c
7	60.0	B	a	B:a
8	61.4	B	a	B:a
9	57.5	B	b	B:b
10	56.9	B	b	B:b
11	61.1	B	c	B:c
12	58.9	B	c	B:c



```r
library(ggplot2)
ggplot(data = Pastes, aes(x = batch, y = strength, col = cask)) +
  geom_point() +
  labs(x = "Batch", y = "Strength", color = "Cask") +
  theme_minimal()
```

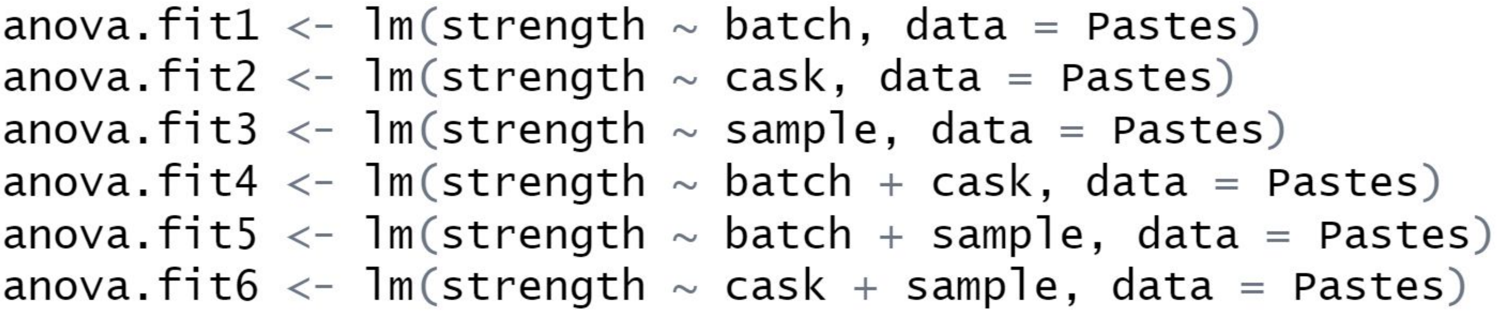
a) Assume you would have forgotten your knowledge on random effects and want to analyse the data set with an one- or two-way ANOVA. How do you have to specify the model, and what are the problems of such an approach?

$$Y_{ij} = \mu + \alpha_i + E_{ij}, \quad E_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

$$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}, \quad E_{ijk} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$$

α_i = batch-, cask- or sample effect

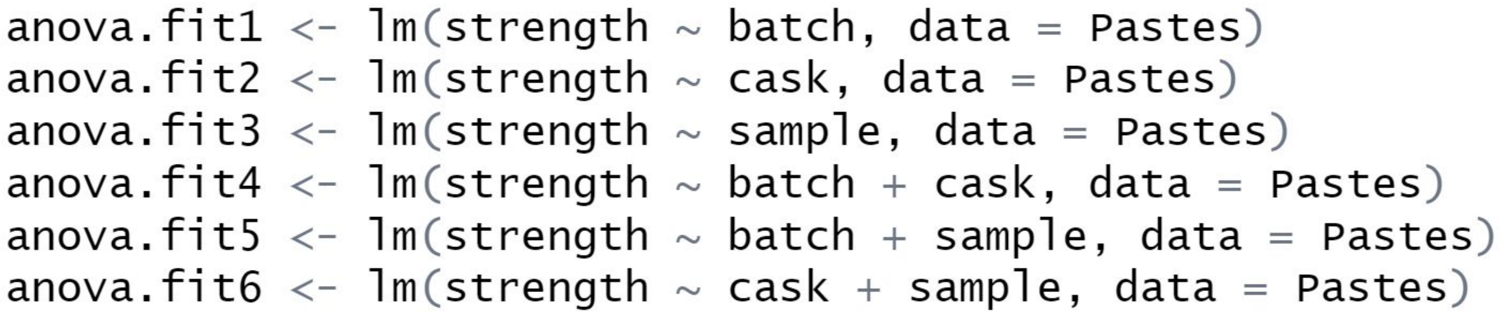
β_i = batch-, cask- or sample effect



```
anova.fit1 <- lm(strength ~ batch, data = Pastes)
anova.fit2 <- lm(strength ~ cask, data = Pastes)
anova.fit3 <- lm(strength ~ sample, data = Pastes)
anova.fit4 <- lm(strength ~ batch + cask, data = Pastes)
anova.fit5 <- lm(strength ~ batch + sample, data = Pastes)
anova.fit6 <- lm(strength ~ cask + sample, data = Pastes)
```

Problems with the approach

- Small group sizes
- Loss of randomness
- Nested design



```r
anova.fit1 <- lm(strength ~ batch, data = Pastes)
anova.fit2 <- lm(strength ~ cask, data = Pastes)
anova.fit3 <- lm(strength ~ sample, data = Pastes)
anova.fit4 <- lm(strength ~ batch + cask, data = Pastes)
anova.fit5 <- lm(strength ~ batch + sample, data = Pastes)
anova.fit6 <- lm(strength ~ cask + sample, data = Pastes)
```

b) Assuming you have not forgotten anything on random effects. Fit a two-way random effects model to the data set. Which model formula is appropriate? Which remarkable feature does the data have?

Random Effects Model

Modelling something as a random effect is appropriate when:

- model a categorical explanatory variable with random cell means
- measurements are not repeatable
- few measurements per explanatory variable
- interest lies more in the variance between experimental conditions than actual effects of a certain condition

Basic Model:

$$Y_{ij} = \mu + \alpha_i + E_{ij} \;\; with \; E_{ij} \; i.i.d. \sim N(0, \sigma^2) \; and \; \alpha_i \; i.i.d. \sim N(0, \sigma^2_{group})$$

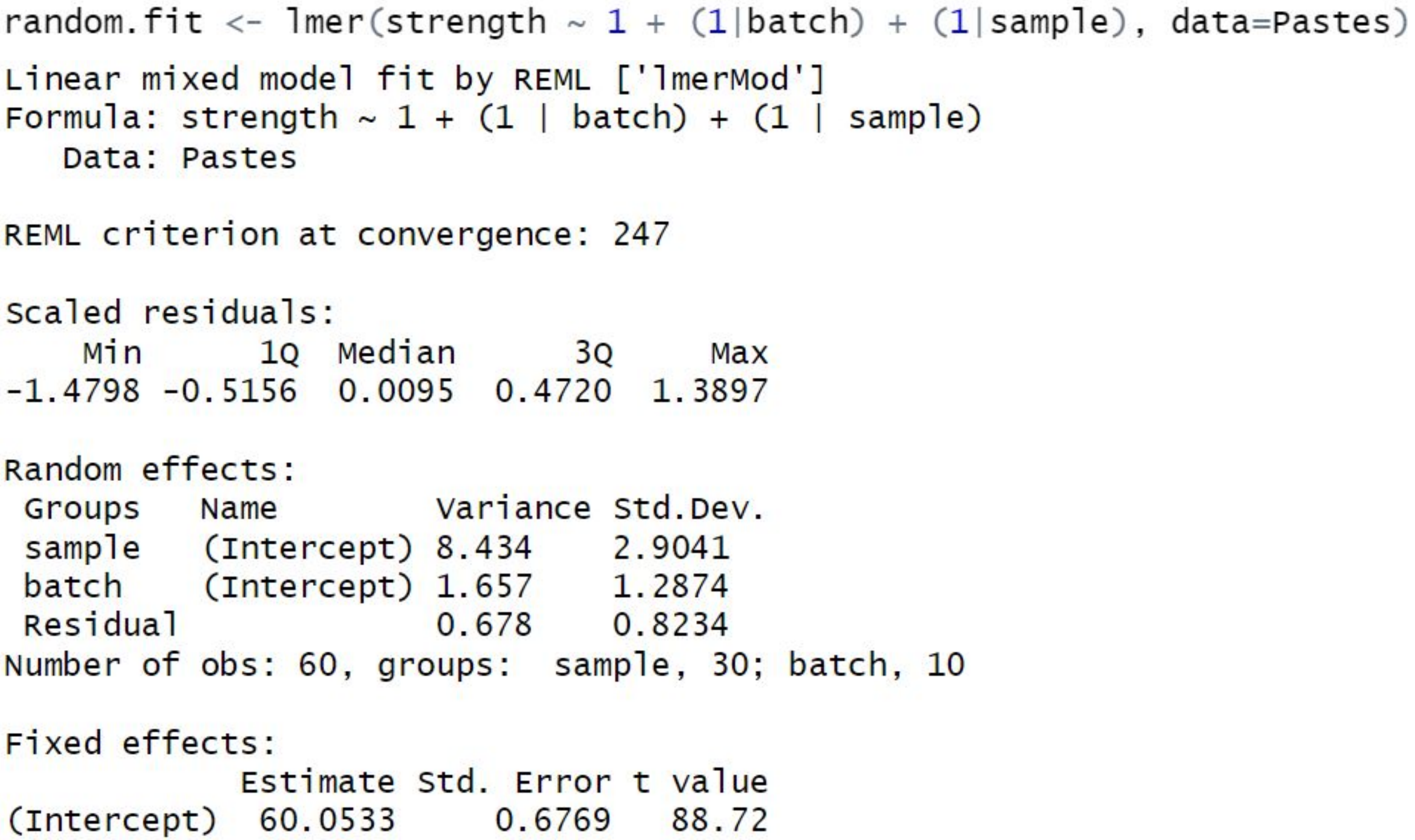
Model

strength ~ batch + sample, both modelled as random effects

$$Y_{ijk} = \mu + a_i + b_{ij} + E_{ijk}$$

$$with\ E_{ijk}\ i.i.d. \sim N(0, \sigma^2), a_i\ i.i.d. \sim N\left(0, \sigma_{batch}^2\right), and\ b_{ij}\ i.i.d. \sim N(0, \sigma_{sample}^2)$$

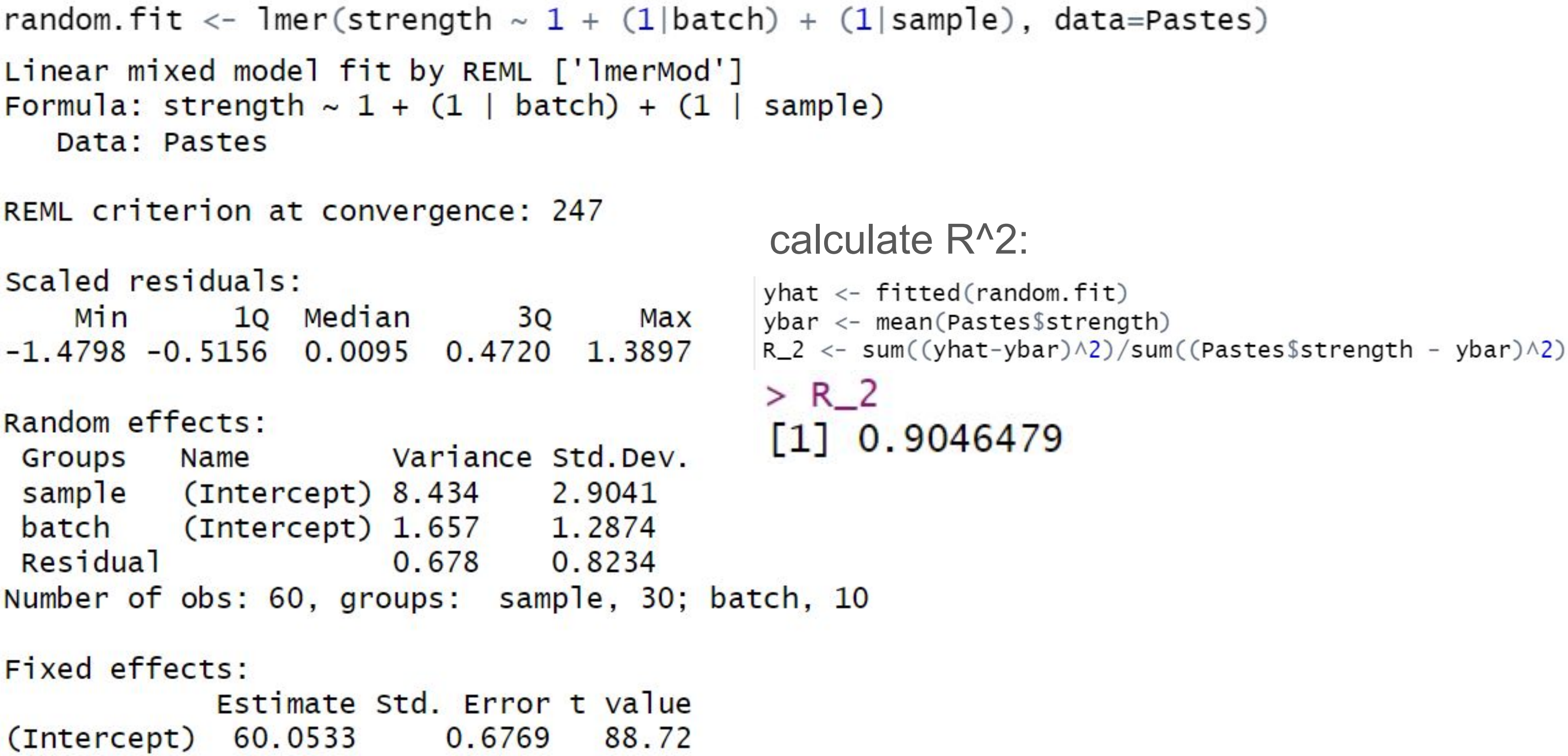
Model in R



```
random.fit <- lmer(strength ~ 1 + (1|batch) + (1|sample), data=Pastes)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: strength ~ 1 + (1 | batch) + (1 | sample)
   Data: Pastes

REML criterion at convergence: 247

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-1.4798 -0.5156  0.0095  0.4720  1.3897 

Random effects:
 Groups   Name        Variance Std.Dev.
 sample   (Intercept) 8.434    2.9041  
 batch    (Intercept) 1.657    1.2874  
 Residual             0.678    0.8234  
Number of obs: 60, groups:  sample, 30; batch, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.0533     0.6769   88.72
```

Model in R



```
random.fit <- lmer(strength ~ 1 + (1|batch) + (1|sample), data=Pastes)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: strength ~ 1 + (1 | batch) + (1 | sample)
   Data: Pastes

REML criterion at convergence: 247

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-1.4798 -0.5156  0.0095  0.4720  1.3897 

Random effects:
 Groups   Name        Variance Std.Dev.
 sample   (Intercept) 8.434    2.9041  
 batch    (Intercept) 1.657    1.2874  
 Residual             0.678    0.8234  
Number of obs: 60, groups:  sample, 30; batch, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.0533     0.6769   88.72
```

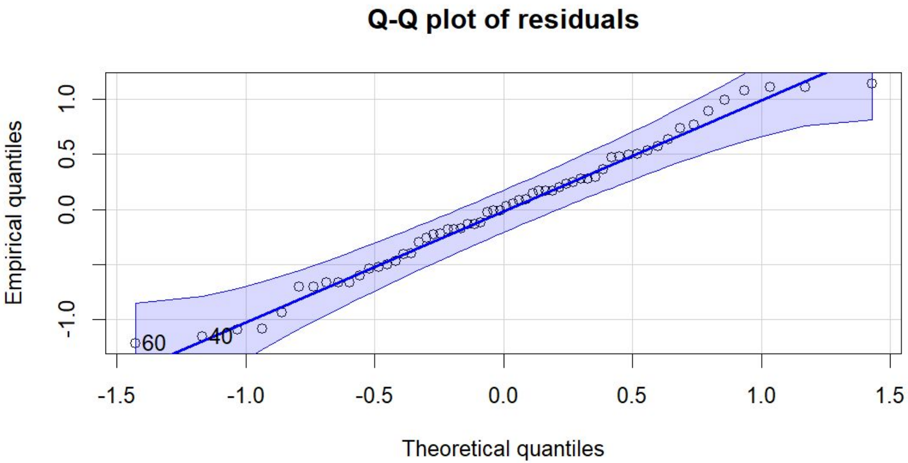
calculate R^2:

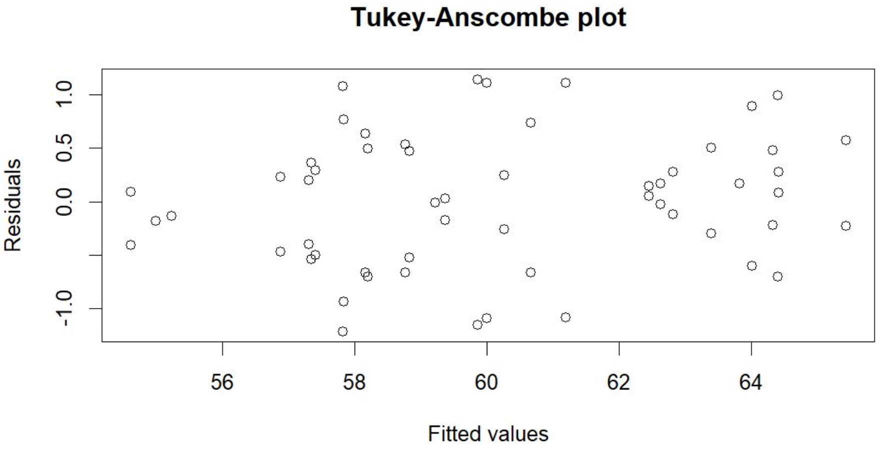
```
yhat <- fitted(random.fit)
ybar <- mean(Pastes$strength)
R_2 <- sum((yhat-ybar)^2)/sum((Pastes$strength - ybar)^2)
```

```
> R_2
[1] 0.9046479
```

Model assumptions

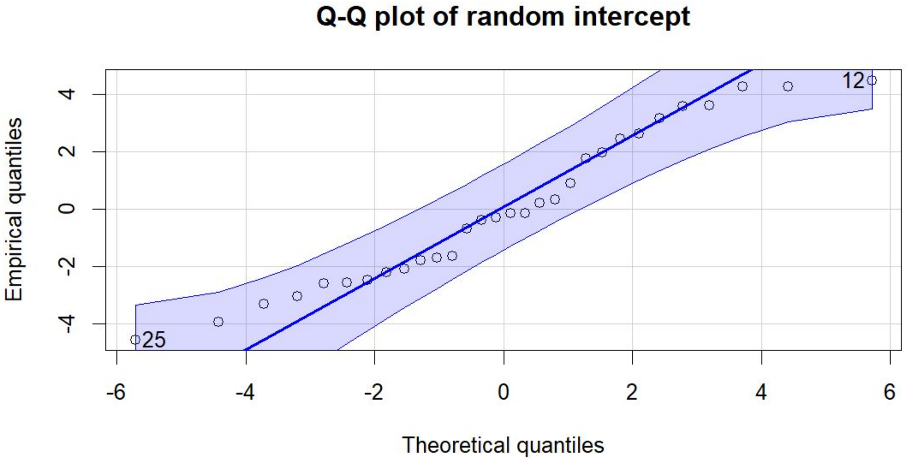
Normality and i.i.d. of error residuals



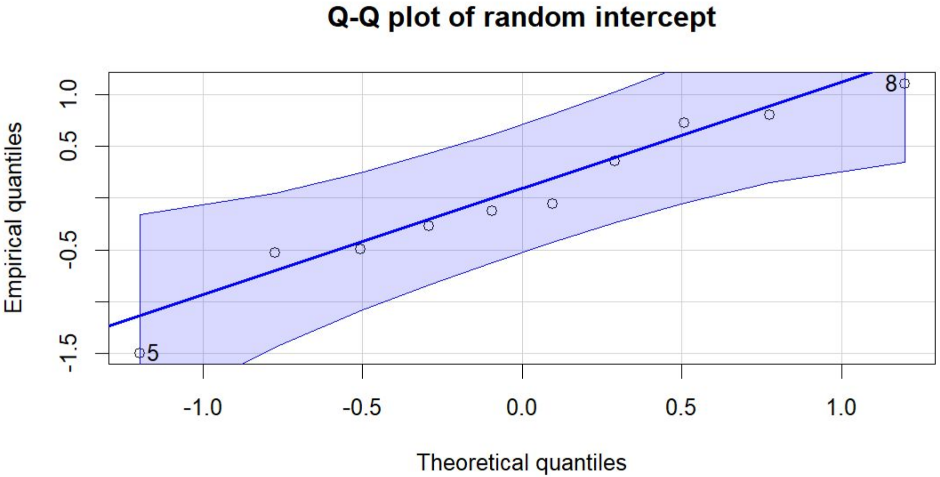


Model assumptions

Normality of sample effect

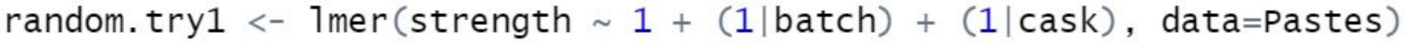


and of batch effect

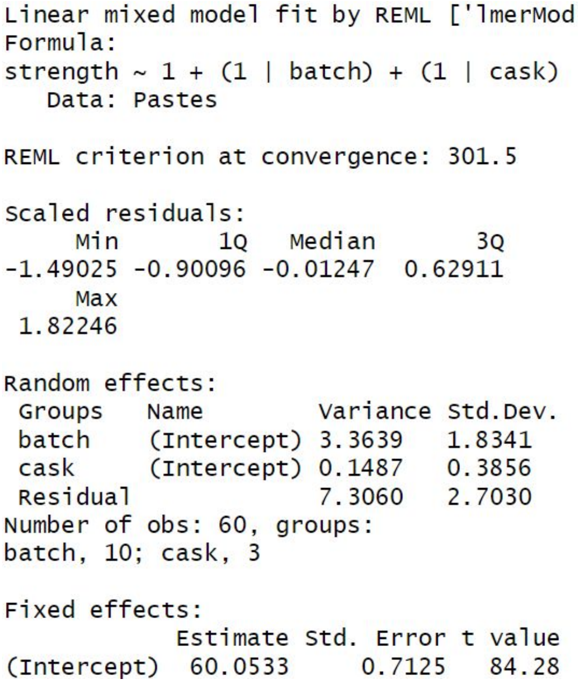


Model in R

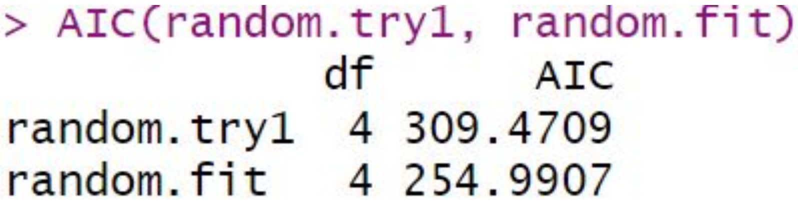
with batch and cask as explanatory variables



```
random.try1 <- lmer(strength ~ 1 + (1|batch) + (1|cask), data=Pastes)
```

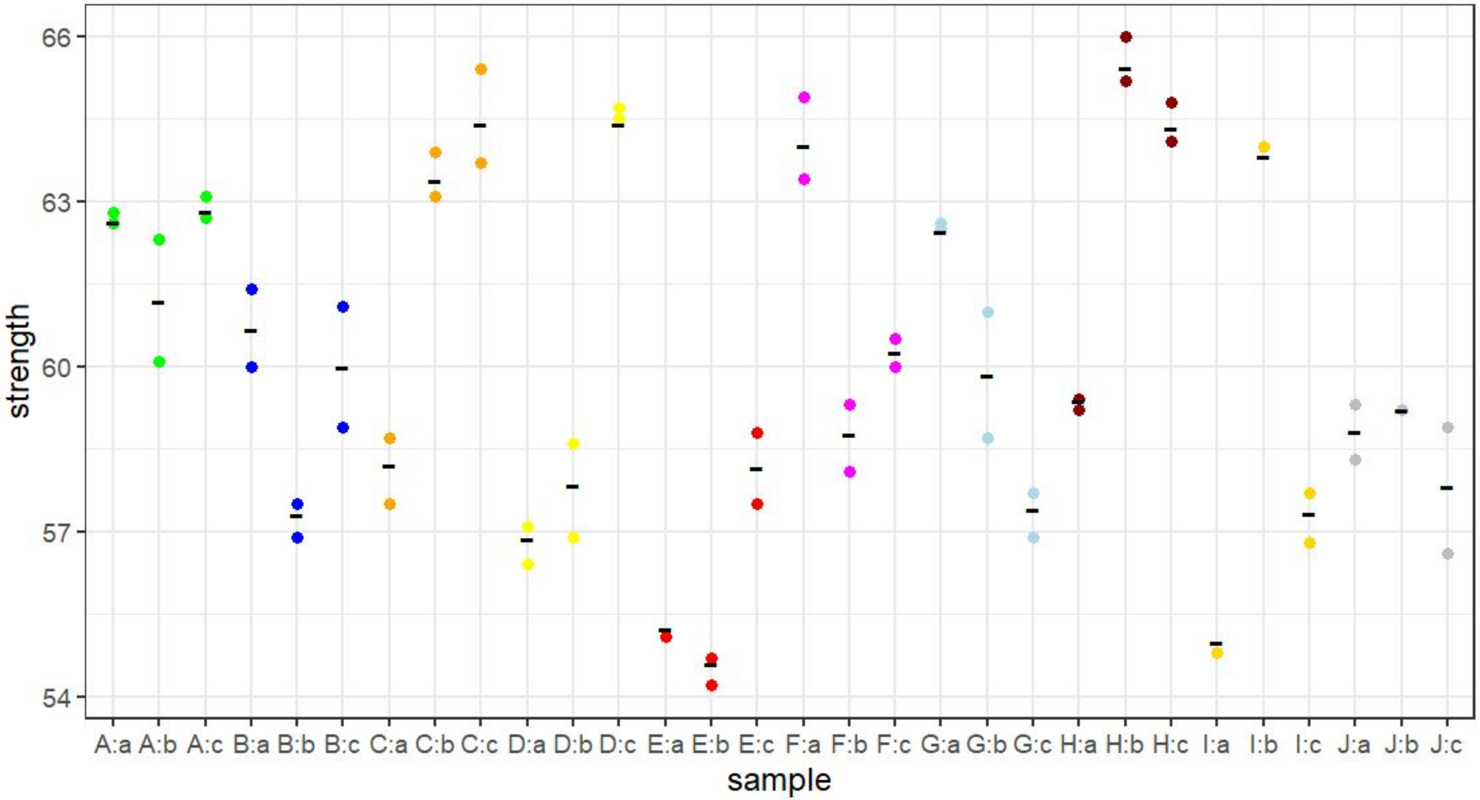


```
Linear mixed model fit by REML ['lmerMod']
Formula:
strength ~ 1 + (1 | batch) + (1 | cask)
   Data: Pastes

REML criterion at convergence: 301.5

Scaled residuals:
     Min       1Q   Median       3Q 
-1.49025 -0.90096 -0.01247  0.62911 
     Max 
 1.82246 

Random effects:
 Groups   Name        Variance Std.Dev.
 batch    (Intercept) 3.3639   1.8341  
 cask     (Intercept) 0.1487   0.3856  
 Residual             7.3060   2.7030  
Number of obs: 60, groups:  
batch, 10; cask, 3

Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.0533     0.7125   84.28
```



```
> AIC(random.try1, random.fit)
            df      AIC
random.try1  4 309.4709
random.fit   4 254.9907
```

AIC shows that we should prefer the model with batch and sample as explanatory variables

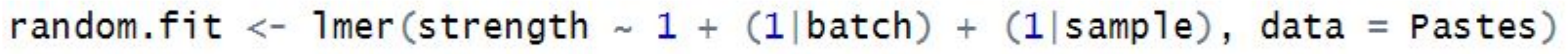

strength
66
63
60
57
54
sample
A:a
A:b
A:c
B:a
B:b
B:c
C:a
C:b
C:c
D:a
D:b
D:c
E:a
E:b
E:c
F:a
F:b
F:c
G:a
G:b
G:c
H:a
H:b
H:c
I:a
I:b
I:c
J:a
J:b
J:c

BOOTSTRAP APPROACH

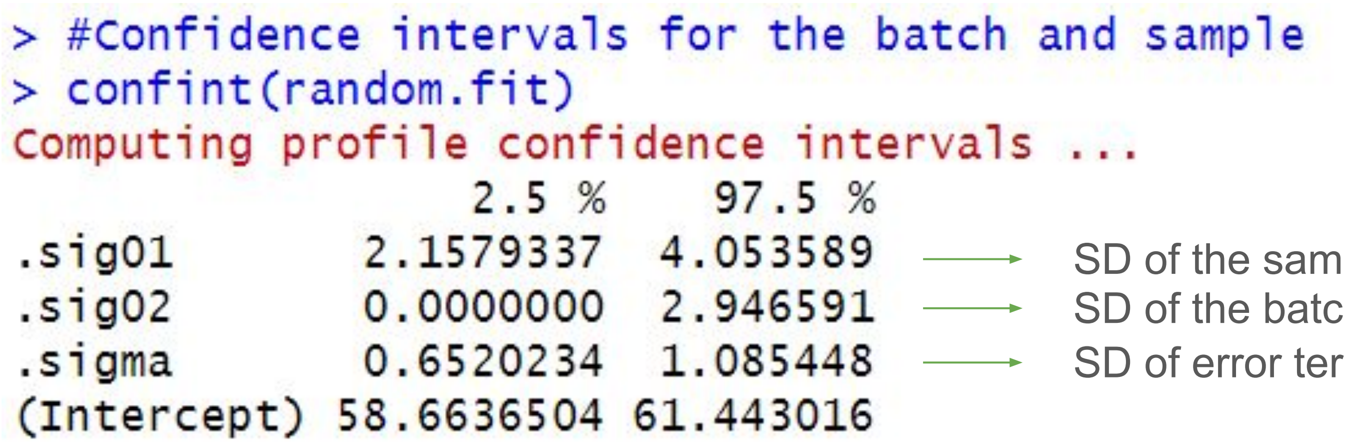
- For ANOVA with random effects there is no test statistic with a known distribution.
- There is no method to calculate an exact p-value as we do in a ANOVA with fixed effects.

- Using resampling to simulate an unknown distribution of an estimator.
    - Estimate the distribution of the data set.
    - Resample from that fitted distribution,(i.e. draw a certain number of samples of the same size as your actual data set from the fitted distribution)
    - Calculate the estimator on all of these (simulated) samples
    - Use the empirical distribution of the estimators as an approximation to the true distribution of the estimator

c) Does the strength significantly deviate between different casks? And **between different batches?**



```
random.fit <- lmer(strength ~ 1 + (1|batch) + (1|sample), data = Pastes)
```



```
> #Confidence intervals for the batch and sample
> confint(random.fit)
Computing profile confidence intervals ...
                 2.5 %    97.5 %
.sig01        2.1579337  4.053589
.sig02        0.0000000  2.946591
.sigma        0.6520234  1.085448
(Intercept) 58.6636504 61.443016
```

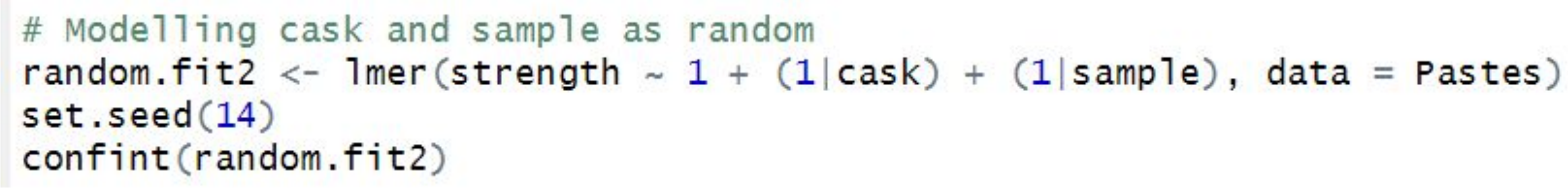
SD of the sample

SD of the batch

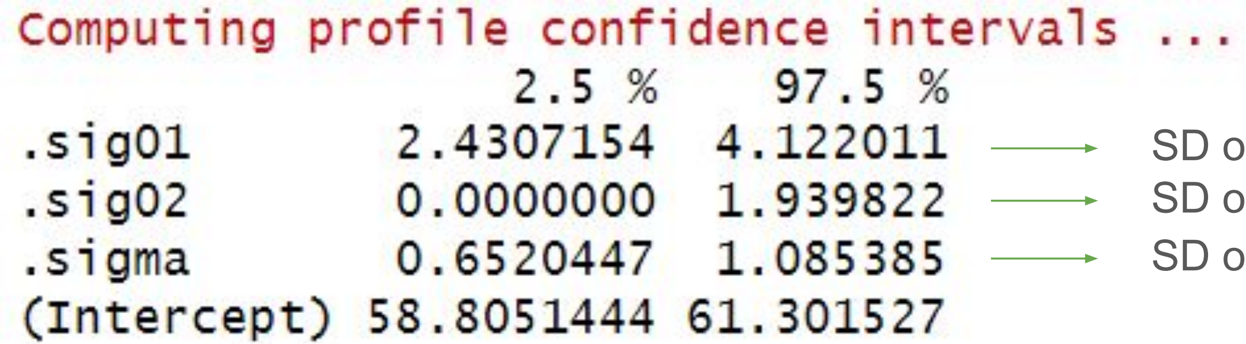
SD of error terms

There is no significant variability in strength between the different batches

c) Does the strength significantly deviate **between different casks?** And between different batches?



```
# Modelling cask and sample as random
random.fit2 <- lmer(strength ~ 1 + (1|cask) + (1|sample), data = Pastes)
set.seed(14)
confint(random.fit2)
```



```
Computing profile confidence intervals ...
                 2.5 %    97.5 %
.sig01       2.4307154  4.122011
.sig02       0.0000000  1.939822
.sigma       0.6520447  1.085385
(Intercept) 58.8051444 61.301527
```

SD of the sample

SD of the cask

SD of error terms

There is no significant variability in strength between the different casks

Any questions … ?

Concluding remarks

- It’s important to recognise when we should use **fixed effects or random effects**.
- Random effects are often more appropriate when we have **too many explanatory variables or groups** and **small group sizes**.
- We are interested in the **variation across the groups** rather than the effect of individual groups.
- The **bootstrap approach** can be used to test the **confidence of a random effect**, due to there is no a reliable test statistic with a known distribution in these cases.