STAA 554 Homework 4

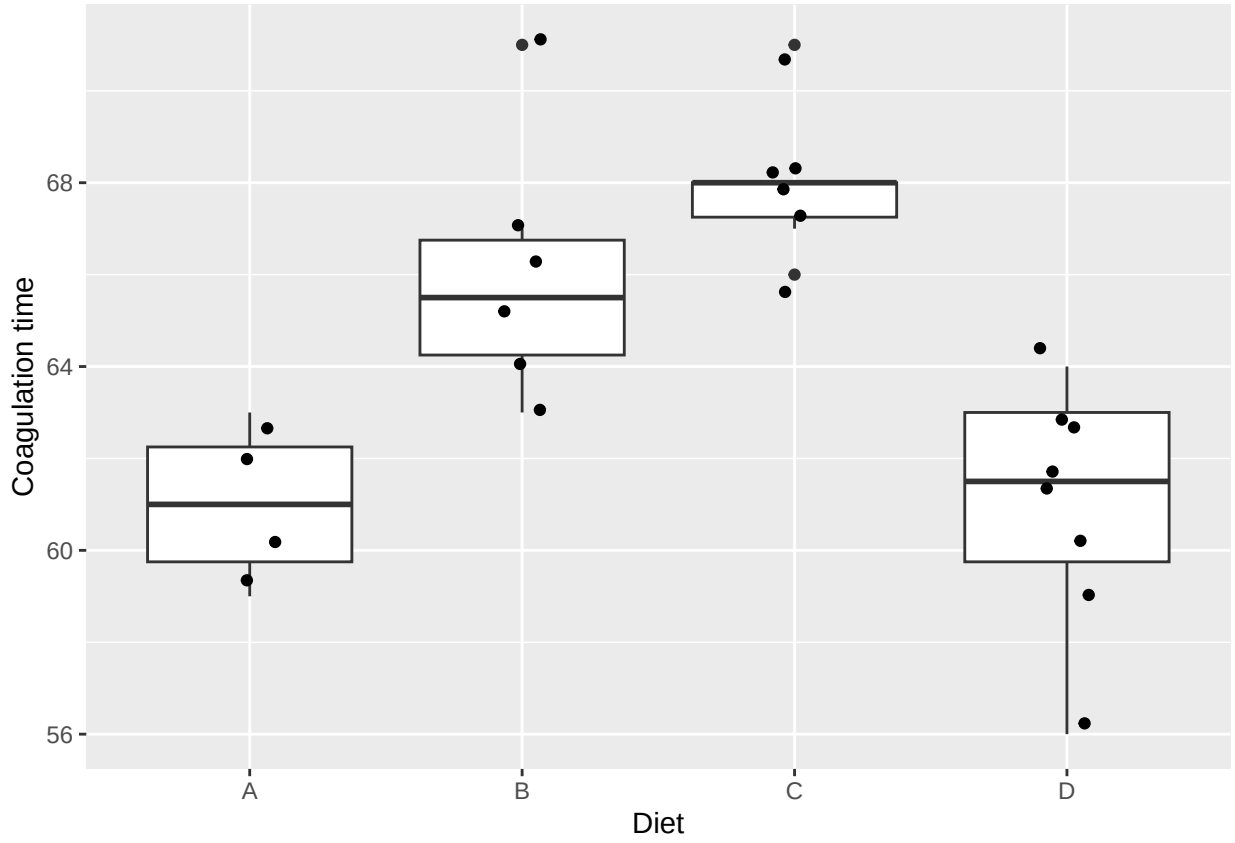
Contents

Q1 Diet and blood coagulation.

The coagulation dataset comes from a study of blood coagulation times. Twenty-four animals were randomly assigned to four different diets and the samples were taken in a random order.

(a) 2pts Plot the data.

```
df1 <- read.csv("Data/coagulation.csv")
ggplot(df1, aes(diet, coag)) +
  geom_boxplot() +
  geom_jitter(width = 0.1) +
  labs(x = "Diet", y = "Coagulation time")
```



(b) 2pts

Fit a fixed effects model and construct a prediction together with a 95% prediction interval for the response of a new animal assigned to diet D.

```
mod1b <- lm(coag ~ diet, data = df1)
new1b <- data.frame(diet = "D")
predict(mod1b, new1b, interval = "prediction", level = 0.95)
```

```
  fit    lwr    upr
1  61 55.764 66.236
```

(c) 4pts

Now fit a random effects model using REML. A new animal is assigned to diet D. Predict the blood coagulation time for this animal along with a 95% prediction interval.

```
mod1c   <- lmer(coag ~ 1 + (1 | diet), data = df1, REML = TRUE)
intcpt   <- fixef(mod1c)[1]
resid_sd <- sigma(mod1c)
diet_sd  <- attr(VarCorr(mod1c)$diet, "stddev")
```

```
diet_re  <- ranef(mod1c)$diet["D", "(Intercept)"]
pred_D   <- intcpt + diet_re
pi_D     <- pred_D + c(lower = -1.96, fit = 0, upper = 1.96) * resid_sd
pi_D
```

```
 lower    fit  upper
56.532 61.170 65.808
```

(d) 4pts

A new diet is given to a new animal. Predict the blood coagulation time for this animal along with a 95% prediction interval

```
pred_Z   <- intcpt
se_Z     <- sqrt(resid_sd^2 + diet_sd^2)
pi_Z     <- pred_Z + c(lower = -1.96, fit = 0, upper = 1.96) * se_Z
pi_Z
```

```
 lower    fit  upper
55.863 64.013 72.163
```

(e) 4pts

A new diet is given to the first animal in the dataset. Predict the blood coagulation time for this animal with a prediction interval. You may assume that the effects of the initial diet for this animal have washed out. (Open for discussion – how to get at the animal effect for this prediction?)

```
pred  <- fixef(mod1c)[1]
se    <- sqrt(sigma(mod1c)^2 + attr(VarCorr(mod1c)$diet,"stddev")^2)
pi    <- pred + c(lower = -1.96, fit = 0, upper = 1.96) * se
round(pi, 2)
```

```
lower   fit upper
55.86 64.01 72.16
```

This Prediction is the same as part d because we are assuming that each animal only has one reading. as such the baseline and residual error are confounded. this means that we dont have a way to isolate the baseline, and have to predict the grand mean with a confidence interval of a new animal. We would need multiple records per animal to really isolate the per animal effect.

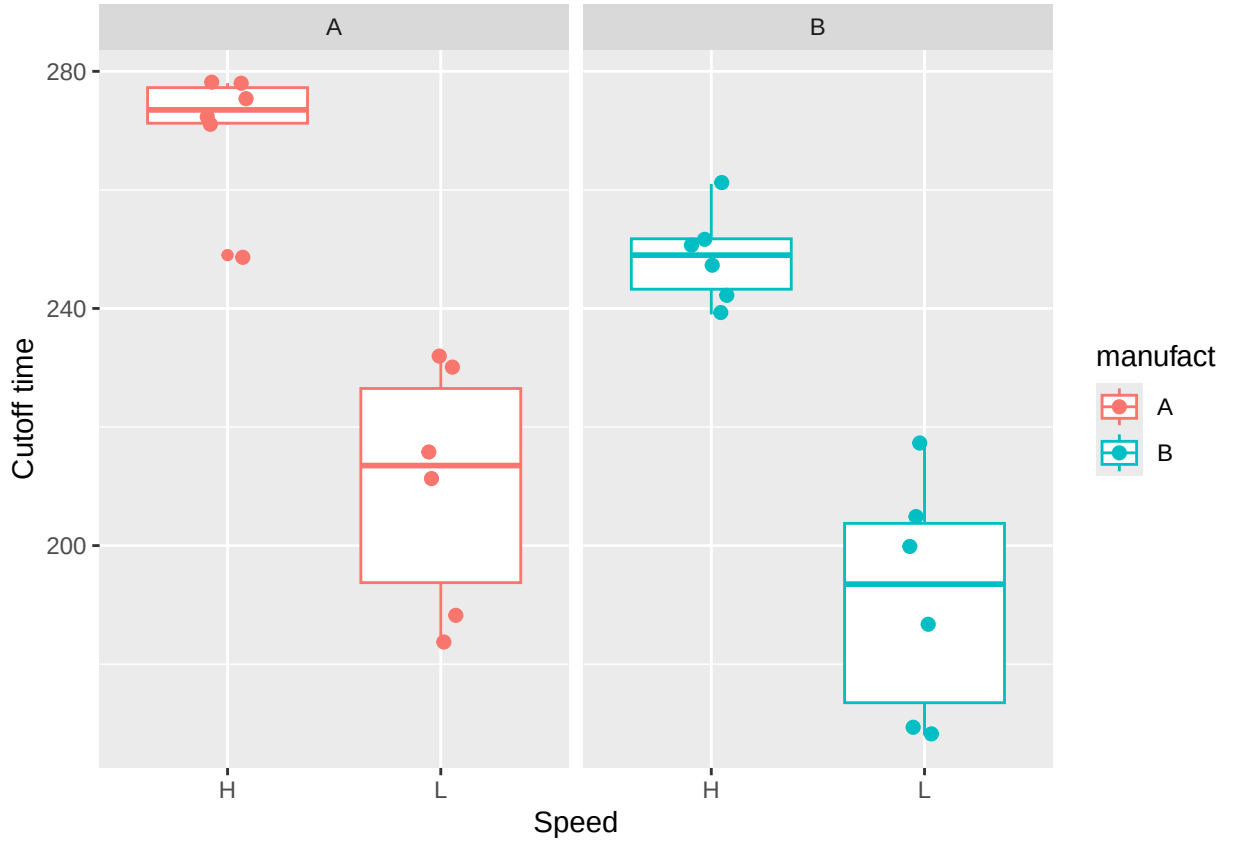
Q2 Lawnmowers and cutoff times.

Data on the cutoff times of lawnmowers may be found in the data set lawn. Three machines were randomly selected from those produced by manufacturers A and B. Each machine was tested twice at low speed and high speed.

(a) 2pts

Make plots of the data and comment.

```
df2 <- read.csv("Data/lawn.csv")

ggplot(df2, aes(speed, time, color = manufact)) +
  geom_boxplot() +
  geom_jitter(width = 0.1, size = 2) +
  facet_wrap(~ manufact) +
  labs(x = "Speed", y = "Cutoff time")
```



Going from Low to High speed adds about 1 minute of cutoff time for both manufacturers. there does appear to be a manufacturer difference as well, especially at high speed, but this effect is smaller around 30-40 seconds.

(b) 2pts

Fit a fixed effects model for the cutoff time response using just the main effects of the three predictors. Explain why not all effects can be estimated.

```
mod2b <- lm(time ~ manufact + machine + speed, data = df2)
summary(mod2b)
```

```
Call:
lm(formula = time ~ manufact + machine + speed, data = df2)

Residuals:
   Min     1Q Median     3Q    Max
```

```
 -12.5   -8.5   -3.0    7.5   19.2

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   278.75       6.04   46.14  < 2e-16
manufactB     -26.75       7.91   -3.38   0.0035
machinem2     -26.00       7.91   -3.29   0.0044
machinem3      -0.75       7.91   -0.09   0.9256
machinem4       7.50       7.91    0.95   0.3563
machinem5     -15.50       7.91   -1.96   0.0667
machinem6         NA         NA      NA       NA
speedL        -59.00       4.57  -12.92  3.2e-10

Residual standard error: 11.2 on 17 degrees of freedom
Multiple R-squared:  0.925, Adjusted R-squared:  0.899
F-statistic:   35 on 6 and 17 DF,  p-value: 1.18e-08
```

Machine is not always estimated because it is nested within manufacturer. when nested factors are included additively, the fixed effect columns for machine add up to the manufacturer level meaning that the design matrix is rank-deficient. to address this we do not get a coefficient for machine6. We would address this by explicitly nesting machine within manufacturer or using random effects

(c) 2pts

Fit a mixed effects model with manufacturer and speed as main effects along with their interaction; include machine as a random effect.

- If the same machine were tested at the same speed, what would be the SD of the times observed?
- If different machines were sampled from the same manufacturer and tested at the same speed once only, what would be the SD of the times observed?

```r
mod2c      <- lmer(time ~ manufact * speed + (1 | machine), data = df2, REML = TRUE)
resid_sd2  <- sigma(mod2c)
mach_sd2   <- sqrt(as.data.frame(VarCorr(mod2c))$vcov[1])
c(residual_sd         = resid_sd2,
  different_machine_sd = sqrt(resid_sd2^2 + mach_sd2^2))
```

```
        residual_sd different_machine_sd
             11.502              16.659
```

- If the machine were tested at the same speed, the standard deviation would be 11.5s
- if different machines were sampled from the same manufacturer, the standard deviation would be 16.659s

(d) 3 pts

Test whether the interaction term of the model can be removed. If so, go on to test the two main fixed effects terms.

```
mod2c_ML  <- lmer(time ~ manufact * speed + (1 | machine),
                  data = df2, REML = FALSE)
mod2d1_ML <- update(mod2c_ML, . ~ manufact + speed + (1 | machine))   # no interaction
mod2d2_ML <- update(mod2d1_ML, . ~ manufact + (1 | machine))          # drop speed
mod2d3_ML <- update(mod2d1_ML, . ~ speed + (1 | machine))             # drop manufact

print("Interaction Term")
```

```
[1] "Interaction Term"
```

```
anova(mod2c_ML, mod2d1_ML)
```

```
Data: df2
Models:
mod2d1_ML: time ~ manufact + speed + (1 | machine)
mod2c_ML: time ~ manufact * speed + (1 | machine)
          npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
mod2d1_ML    5 201 207  -95.5      191
mod2c_ML     6 203 210  -95.5      191  0.09  1       0.76
```

```
print("Speed as Main effect")
```

```
[1] "Speed as Main effect"
```

```
anova(mod2d1_ML, mod2d2_ML)
```

```
Data: df2
Models:
mod2d2_ML: time ~ manufact + (1 | machine)
mod2d1_ML: time ~ manufact + speed + (1 | machine)
          npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
mod2d2_ML    4 244 248 -117.8      236
mod2d1_ML    5 201 207  -95.5      191  44.7  1    2.3e-11
```

```
print("Manufacturer as main effect")
```

```
[1] "Manufacturer as main effect"
```

```
anova(mod2d1_ML, mod2d3_ML)
```

```
Data: df2
Models:
mod2d3_ML: time ~ speed + (1 | machine)
mod2d1_ML: time ~ manufact + speed + (1 | machine)
          npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
mod2d3_ML    4 203 208  -97.4      195
mod2d1_ML    5 201 207  -95.5      191   3.8  1      0.051
```

We see that the interaction can be dropped with a p value of .76. Speed maters a lot both as seen in our plots and here with a very small p value, and the machine fixed effect is significant at a p = .1 level but borderline not signifficant at a p=.051 level. In this case, I would retain manufacturer as it is borderline significant. This makes my final recommended model to be time ~ speed + manufacturer + (1 | machine). Additionally AIC prefers retaining manufacturer by a small margine while BIC does not.

(e) 2pts

Test whether there is any variation between machines. Use the model from part (d) with non-significant terms removed.

```
m_with <- lmer(time ~ speed + manufact + (1 | machine),
               data = df2, REML = FALSE)

rand(m_with)
```

```
ANOVA-like table for random-effects: Single term deletions

Model:
time ~ speed + manufact + (1 | machine)
              npar logLik AIC  LRT Df Pr(>Chisq)
<none>           5  -95.5 201
(1 | machine)    4  -98.1 204 5.16  1      0.023
```

There is significant variance between machines at a .05 level as shown above (p = .023).

(f) 2pts

Fit a model with speed as the only fixed effect and manufacturer as a random effect with machines also as a random effect nested within manufacturer. Compare the variability between machines with the variability between manufacturers.

```
mod2f <- lmer(time ~ speed + (1 | manufact/machine), data = df2, REML = TRUE)

vc <- as.data.frame(VarCorr(mod2f))

var_manuf <- vc$vcov[vc$grp == "manufact"]
var_mach  <- vc$vcov[vc$grp == "machine:manufact"]
var_resid <- vc$vcov[vc$grp == "Residual"]

tot_var   <- var_manuf + var_mach + var_resid

pct_manuf <- 100 * var_manuf / tot_var
pct_mach  <- 100 * var_mach  / tot_var
pct_resid <- 100 * var_resid / tot_var

out <- data.frame(
  source    = c("manufacturer", "machine:manufacturer", "residual"),
  variance  = c(var_manuf, var_mach, var_resid),
  percent   = c(pct_manuf, pct_mach, pct_resid)
)
```

```
out
```

```
                source variance percent
1         manufacturer   150.69  35.637
2 machine:manufacturer   147.02  34.767
3             residual   125.15  29.596
```

Manufacturer-manufacturer variability is very similar to machine-machine variability. when we look at total unexplained variance, manufacturers account for 35.6%, machine:manufacturer for 34.8% and residual variance is 29.6%.

(g) 2pts

Construct bootstrap confidence intervals for the terms of the previous model. Discuss whether the variability can be ascribed solely to manufacturers or to machines. *hint: the confint() command allows a variety of method options.*

```
set.seed(420)
confint(mod2f, method = "boot", nsim = 1000, parm = "theta_")
```

```
       2.5 % 97.5 %
.sig01  0.00 20.054
.sig02  0.00 33.277
.sigma  7.25 14.717
```

Because the first two intervals include zero, the amount of data we have is too small to exclude the possibility that either the manufacturer of machine differences are zero. Our SD estimates suggest that if the effects are real, the variability is approximately equal but we cannot attribute this to one of the other statistically.

Q3 5 pts Normality assumption of random effects

At the end of the School Performance Lab, there is a suggested way to explore the robustness of the normality assumption. Choose one of the following:

- the suggested gamma distribution
- the suggested double exponential distribution
- or a distribution of your choice

Then:

```
1. simulate new values for the underlying or performance of each school ("mu").
This is done in section "1. Generate true performance..."
2.  generate new data from these "mus" as in section "2. Generate observed...".
Save this data to a data set to be used in the following step.
3. Fit the empirical Bayes estimates as in section "Empirical Bayes estimates..."
```

\vskip .1in **Include your results HERE in your submission, and ALSO on Canvas Discussion Board.**
\vskip .1in
Summarize your results on the discussion board under either the "Gamma simulation" or the "Double exponential simulation", or "Other simulation" conversation as appropriate. In one post please include:

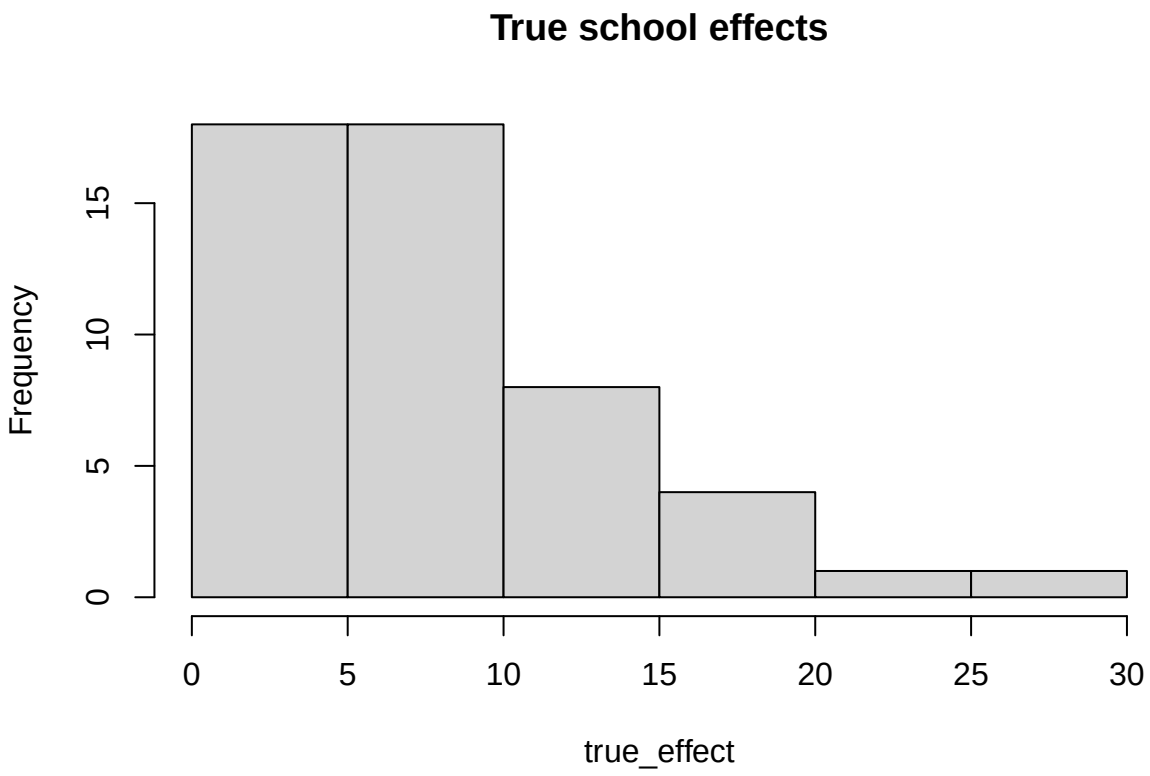
a. a histogram of your simulated values.

b. a summary of the error in the random effect estimates, for example:

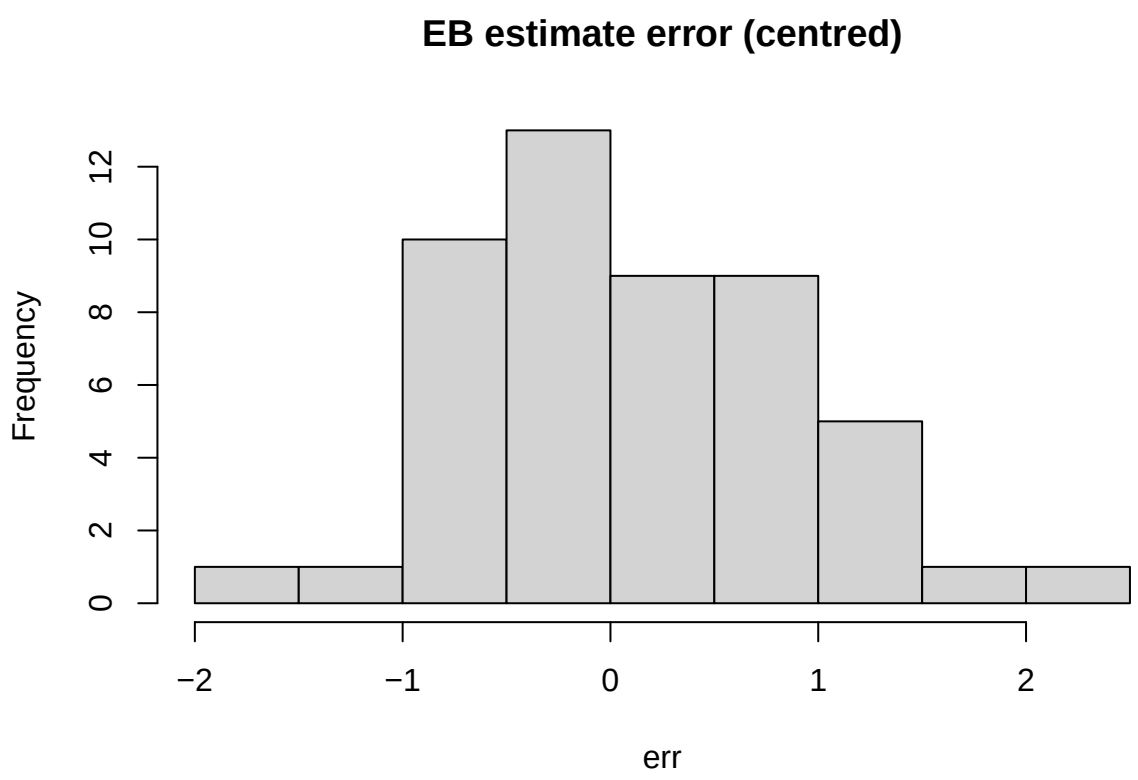
```
hist(coef(mod)$schools[,1] - sorted.data$mu)
sd(coef(mod)$schools[,1] - sorted.data$mu)
```

c. Report the estimated school variance component, and the true value. (Please use standard deviation.)

```r
set.seed(123)
n_schools      <- 50
obs_per_school <- 10
true_effect    <- rgamma(n_schools, shape = 2, scale = 5)

sim <- data.frame(
  school = rep(paste0("S", 1:n_schools), each = obs_per_school),
  y      = unlist(lapply(true_effect,
                         function(m) rnorm(obs_per_school, mean = m, sd = 3)))
)

sim$school <- factor(sim$school, levels = paste0("S", 1:n_schools))
mod3 <- lmer(y ~ 1 + (1 | school), data = sim, REML = TRUE)

# a
hist(true_effect, main = "True school effects")
```




```
# b
beta0 <- fixef(mod3)[1]
err   <- (beta0 + ranef(mod3)$school[,1]) - true_effect
hist(err, main = "EB estimate error (centred)")
```




```
sd(err)
```

```
[1] 0.77167
```

```
# c


est_sd  <- sqrt(as.data.frame(VarCorr(mod3))$vcov[1])
true_sd <- sd(true_effect)
c(estimated_school_sd = est_sd,
  true_school_sd      = true_sd)
```

```
estimated_school_sd      true_school_sd
             5.7838              5.6778
```