Homework I

J Lacasa

2026-02-03

1 Statistics in research

In your own words, describe why you need statistical models in your research project.

Answer

Any variation of "I need statistics because I can't observe my variable of interest, so I need to estimate it. To estimate my variable of interest", etc.

2 General linear model

Using mathematical notation, describe the most common statistical model you could think of to describe continuous data.

Answer

Keep in mind that we are describing continuous data where those "continuous data" are the response. The predictors could be continuous or categorical. This model can be typically written out as

$$y_i \sim N(\mu_i, \sigma^2),$$

where y_i is the ith observation of the response, μ_i is the expected value, and σ^2 is the variance for that observation.

2.1 General linear model assumptions

What are the assumptions behind this model?

Answer

Linearity (typically), normality, independence, constant variance.

2.2 How can those assumptions be relaxed?

Answer

- Linearity (typically), by changing the predictor.
- Normality, by changing the distribution of the data.
- Independence, by including random effects.
- Constant variance, by changing the distribution or modeling the variance.

3 In the context of mixed-effects statistical models:

3.1 Fixed vs. random effects

Mention what you think are the most important characteristics/differences between fixed effects vs. random effects.

	Fixed effect	**Random effect**
Where does it go (marginal distribution)	Expected value	Variance-covariance
Inference	Constant for all groups in the population of study	Differ from group to group
Usually used to model	Carefully selected treatments or genotypes	The study design (aka structure in the data, or what is similar to what)
Assumptions	$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1})$	$\boldsymbol{u} \sim N(\mathbf{0}, \mathbf{G})$
Method of estimation	Maximum likelihood, least squares	Restricted maximum likelihood (shrinkage)

3.2 Fixed vs. random effects – cont.

In your own words, explain to the applied user what you think are three key factors to decide whether to model an effect as fixed or random.

Answer:

In designed experiments: Typically, you want your treatment factors to act as a fixed effect. this is because you spent a lot of time determining those treatment levels and thus, they are not random but been carefully selected.

Beyond designed experiments, one must consider what is estimated. Fixed effects estimate treatment effects while random effects estimate variance components. Then, variance components are hardly estimated with only a few levels.

4 Statistical modeling

Write down the statistical models for the following data using mathematical notation.

4.1 Data generated by a designed experiment

Answer:

A designed experiment studying the effects of plant density (low, medium, high) and genotype (5 genotypes) effects with a 3×5 factorial treatment structure and a split-plot design in an RCBD, where plant density (2 levels) is at the whole-plot level, and the genotypes are at the spli-plot level.

The model can be written out as

$$y_{ijk}|\boldsymbol{u} \sim P(\mu_{ijk}, \psi), g(\mu_{ijk}) = \eta_0 + PD_i + G_j + (PD \times G)_{ij} + b_k + wp_{i(k)},$$

where y_{ijk} is the observation from the ith plant density, jth genotype and kth block, μ_{ijk} is the expected value of y_{ijk}, σ^2 is the residual variance, ψ is the dispersion of y_{ijk}, $g(\cdot)$ is the link function, η_0 is the overall mean of the linear predictor, PD_i is the effect of the ith plant density, G_j is the effect of the jth genotype, $(PD \times G)_{ij}$ is the effect of the interaction between the ith plant density and the jth genotype, b_k is the effect of the kth block, $wp_{i(k)}$ is the whole plot effect of the ith PD in the kth block, and $wp_{i(k)} \sim N(0, \sigma^2_{wp})$.

4.2 Opportunistic data

An opportunistic set of data that is the combination of multiple variety trials. The dataset contains information from 15 years, 60 locations (not always does a year include all 60 locations), and over 200 genotypes that are inconsistently present across locations (i.e., not always does a site-year include all genotypes). The objective thus far is to model yield as a function of years, location, and genotype, with a focus on studying the variability across years, locations, and genotypes.

Answer:

Note that the focus is put on studying the variability across years, locations, and genotypes.

Then, the model

$$y_{ijk}|\boldsymbol{u} \sim N(\mu_{ijk}, \sigma^2_\varepsilon), \mu_{ijk} = \mu_0 + G_i + Y_j + L_k + GY_{ij} + GL_{ik} + YL_{jk}, G_i \sim N(0, \sigma^2_G), Y_j \sim N(0, \sigma^2_Y), L_k \sim N(0, \sigma^2_L), GY_{ij} \sim N(0,$$

where y_{ijk} is the observed yield of the ith genotype in the kth location and kth year, G_i is the (random) effect of the ith genotype, Y_k is the effect of the kth location, L_k is the effect of the kth year, GY_{ij}, GL_{ik}, and YL_{jk} are the crossed random effects. Then, the variance components σ^2_m directly quantify the variability across genotypes, years, and locations. Investigators can make inference directly out of those variance components.

4.3 Observational data

In this last example, the data correspond to observations of the native species in a given place. The data were collected in 40 points in a region in the Argentinean Patagonian Steppe (also known as Patagonian Desert) by a group of researchers. They collected 40 points in approximately random locations in that region for 25 years. At each site, the researchers randomly tossed a $1m^2$ loop and counted the total number of exotic (i.e., non-native species) plants wherever it fell. The researchers registered the total number of plants and the total number of exotic plants. They are interested in the overall proportion of exotic plants, and whether it has been increasing in the past decades. One can assume that the overall number of total plants per m^2 is approximately constant in all the points.

Answer:

The model can be decribed generally as

$$y_{ij} \sim \text{Binomial}(\pi_{ij}, N_{ij}), \text{logit}(\pi_{ij}) = \beta_0 + \beta_1 \cdot year_{ij},$$

where y_{ij} is the number of exotic plants found in the ith toss on the jth year of the study, π_{ij} is the probability of finding exotic plants found in the ith toss on the jth year of the study, N_{ij} is the total number of plants in the 1 m^2 loop, β_0 is the intercept in the linear predictor, β_1 is the slope (i.e., the expected increase in the linear predictor for an increase in one year) , $year_{ij}$ is the year.

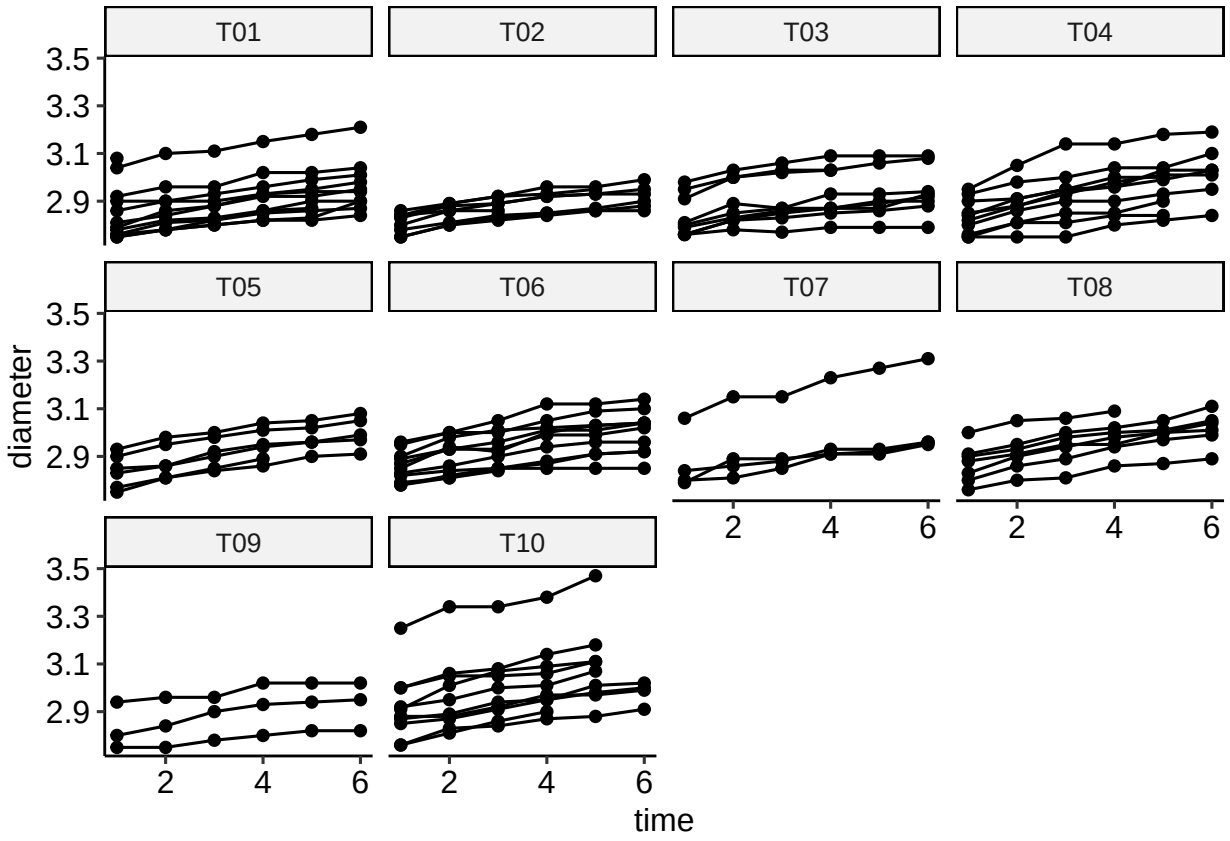
5 Write down the statistical model for the following data, and fit said model using R.

The data below were generated by an experiment that was run to study the growth of apples (in diameter). The experiment conducted at the Winchester Agricultural Experiment Station of Virginia Polytechnic Institute and State University. 25 apples were chosen from each of ten apple trees. The diameters of the apples were recorded every two weeks over a 12-week period.

Answer:

```
library(tidyverse)
library(ggpubr)
```

```
dat <- agridat::byers.apple |>
  drop_na(diameter)
dat |>
  ggplot(aes(time, diameter))+
  geom_line(aes(group=appleid))+
  geom_point()+
  facet_wrap(~tree)+
  theme_pubr()
```



The model can be written as

$$y_{ijk}|\boldsymbol{u} \sim N(\mu_{ijk}, \sigma^2), \mu_{ijk} = \beta_{0ij} + \beta_{1ij} \cdot t_{ijk},$$

where y_{ijk} is the observation of the ith apple in the jth tree at time k, $\boldsymbol{u}$ is the vector of the random effects, μ_{ijk} is the expected value of y_{ijk}, and σ^2 is the variance. The mean is the sum of the apple-tree-specific intercept $\beta_{0ij} = \beta_0 + b_{00i(j)} + b_{0j}$, where $b_{00i(j)} \sim N(0, \sigma^2_{b0}),\ b_{0i(j)} \sim N(0, \sigma^2_{b00})$ and the apple-tree-specific slope $\beta_{1ij} = \beta_1 + b_{11i(j)} + b_{1j}$, where $b_{11i(j)} \sim N(0, \sigma^2_{b1}),\ b_{1i(j)} \sim N(0, \sigma^2_{b11})$, and t_{ijk} is the time.

5.1 Growth rate

What is the growth rate of the diameter for a single apple of unknown tree? Include a 95% confidence interval.

Answer:

The key here is the "unkown tree" part. We thus want a model that describes apple diameter as a function of time.

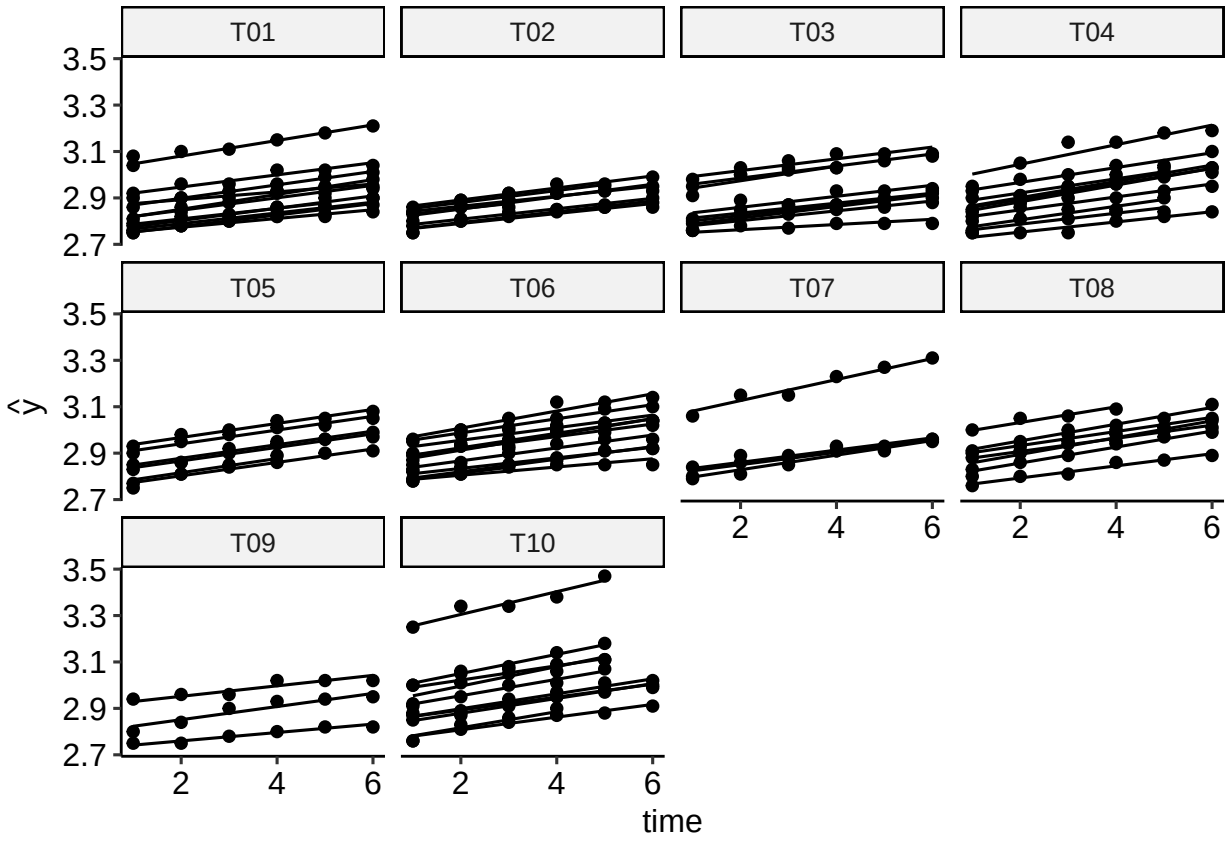
```r
library(lme4)
library(emmeans)

# m <- glmmTMB(diameter ~ 1 + time + ar1(0+factor(time)|tree/apple), data = dat, REML = T)

m <- lmer(diameter ~ 1 + time + (1+time|tree/apple), data = dat)

dat$yhat <- predict(m)

dat |>
  ggplot(aes(time, diameter))+
  geom_line(aes(y = yhat, group=appleid))+
  geom_point()+
  facet_wrap(~tree)+
  labs(y = latex2exp::TeX("$\\hat{y}$"))+
  theme_pubr()
```




```r
emtrends(m, ~ 1, "time")
```

```
##  1       time.trend      SE   df lower.CL upper.CL
##  overall     0.0285 0.00153 8.77   0.0251    0.032
##
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
```

5.2 Compare the intra-tree and the inter-tree variability in growth rate.

As seen below in both the estimated variance components and the distribution of the random effects, the variance component for the apples (i.e., apple × tree) is 0.0062 and larger than the variance component for the trees.

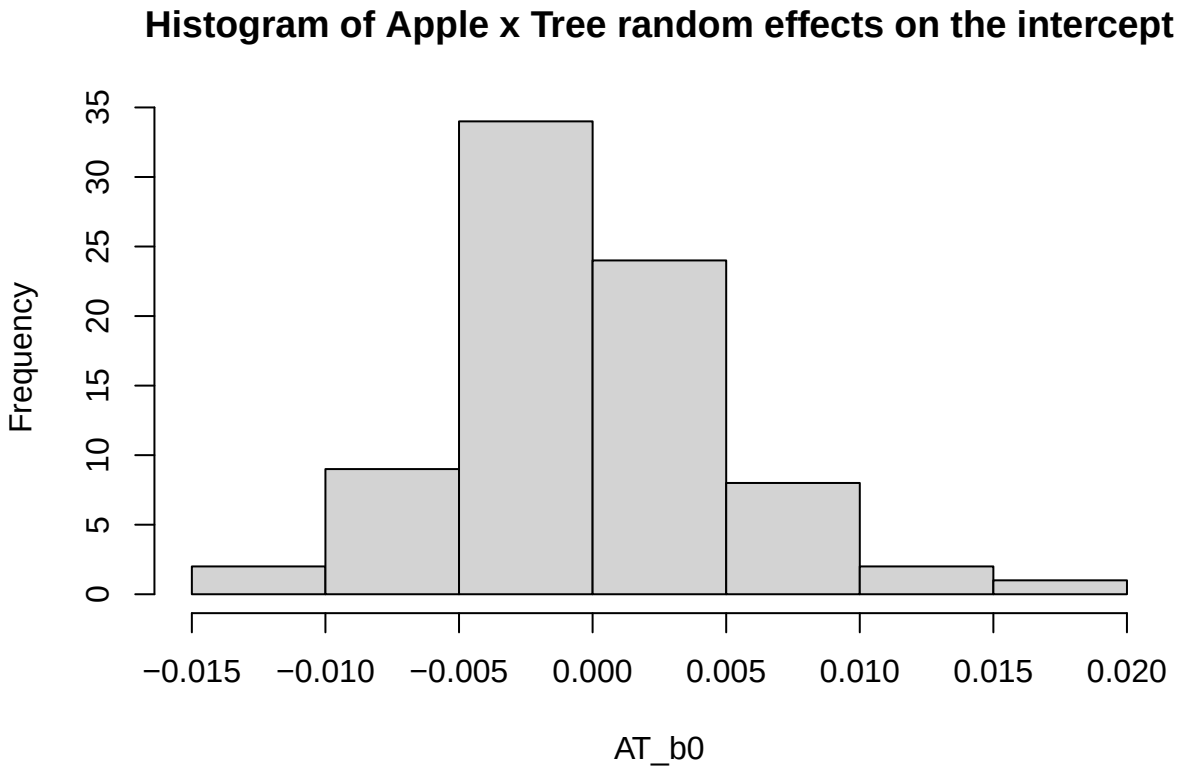
```
VarCorr(m)
```

```
##  Groups     Name        Std.Dev.  Corr
##  apple:tree (Intercept) 0.0882264
##             time        0.0062279 0.533
##  tree       (Intercept) 0.0140447
##             time        0.0039084 1.000
##  Residual               0.0161614
```

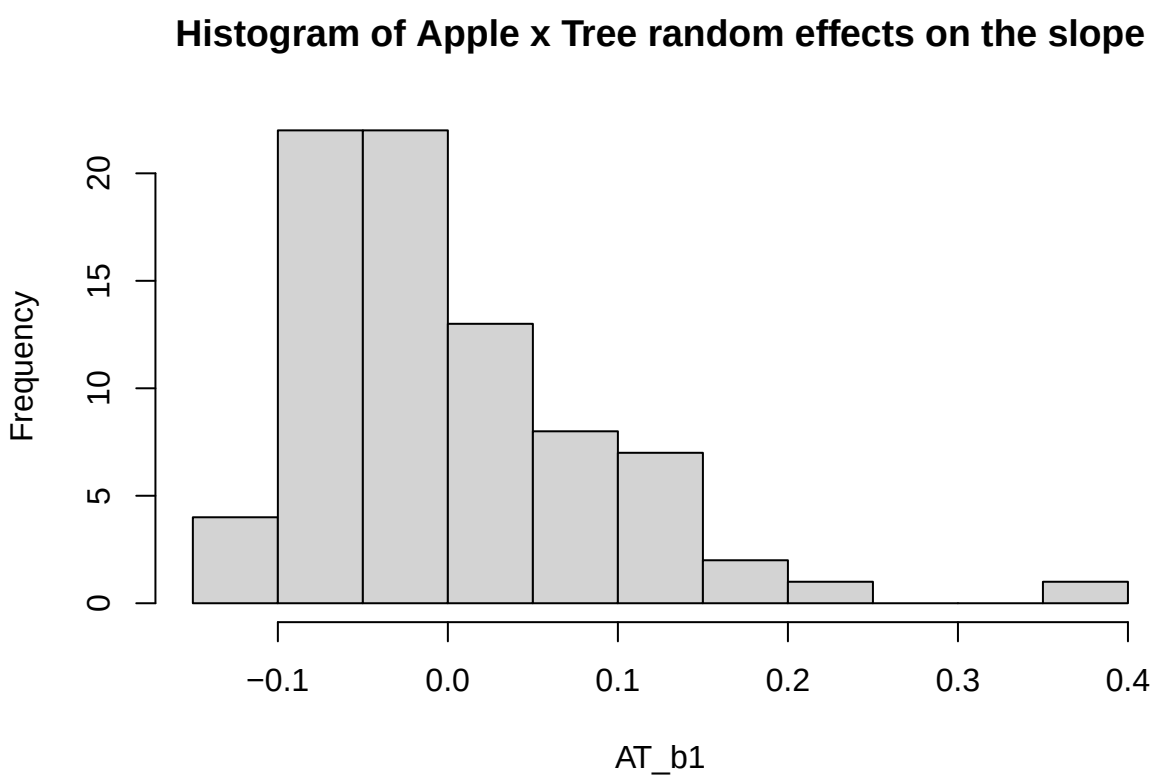
```
colnames(getME(m, "Z"))
```

```
##   [1] "A01:T01" "A01:T01" "A01:T03" "A01:T03" "A01:T04" "A01:T04" "A01:T05"
##   [8] "A01:T05" "A02:T04" "A02:T04" "A02:T07" "A02:T07" "A02:T08" "A02:T08"
##  [15] "A02:T10" "A02:T10" "A03:T04" "A03:T04" "A03:T06" "A03:T06" "A04:T01"
##  [22] "A04:T01" "A04:T06" "A04:T06" "A04:T07" "A04:T07" "A05:T01" "A05:T01"
##  [29] "A05:T06" "A05:T06" "A05:T08" "A05:T08" "A05:T10" "A05:T10" "A07:T02"
##  [36] "A07:T02" "A07:T06" "A07:T06" "A08:T04" "A08:T04" "A08:T09" "A08:T09"
##  [43] "A08:T10" "A08:T10" "A09:T02" "A09:T02" "A09:T07" "A09:T07" "A09:T08"
##  [50] "A09:T08" "A09:T10" "A09:T10" "A10:T01" "A10:T01" "A10:T03" "A10:T03"
##  [57] "A10:T06" "A10:T06" "A10:T09" "A10:T09" "A10:T10" "A10:T10" "A11:T01"
##  [64] "A11:T01" "A11:T02" "A11:T02" "A11:T04" "A11:T04" "A11:T05" "A11:T05"
##  [71] "A12:T06" "A12:T06" "A12:T08" "A12:T08" "A12:T09" "A12:T09" "A13:T01"
##  [78] "A13:T01" "A14:T01" "A14:T01" "A14:T05" "A14:T05" "A14:T06" "A14:T06"
##  [85] "A15:T01" "A15:T01" "A15:T02" "A15:T02" "A15:T08" "A15:T08" "A16:T03"
##  [92] "A16:T03" "A17:T01" "A17:T01" "A17:T02" "A17:T02" "A17:T03" "A17:T03"
##  [99] "A17:T04" "A17:T04" "A17:T06" "A17:T06" "A17:T10" "A17:T10" "A18:T01"
## [106] "A18:T01" "A18:T03" "A18:T03" "A18:T04" "A18:T04" "A18:T05" "A18:T05"
## [113] "A18:T10" "A18:T10" "A19:T01" "A19:T01" "A19:T06" "A19:T06" "A20:T03"
## [120] "A20:T03" "A20:T04" "A20:T04" "A20:T05" "A20:T05" "A20:T06" "A20:T06"
## [127] "A20:T08" "A20:T08" "A21:T08" "A21:T08" "A21:T10" "A21:T10" "A22:T03"
## [134] "A22:T03" "A22:T10" "A22:T10" "A23:T02" "A23:T02" "A23:T03" "A23:T03"
## [141] "A23:T10" "A23:T10" "A24:T02" "A24:T02" "A24:T03" "A24:T03" "A24:T04"
## [148] "A24:T04" "A24:T05" "A24:T05" "A24:T06" "A24:T06" "A25:T01" "A25:T01"
## [155] "A25:T02" "A25:T02" "A25:T04" "A25:T04" "A25:T07" "A25:T07" "T01"
## [162] "T01"     "T02"     "T02"     "T03"     "T03"     "T04"     "T04"
## [169] "T05"     "T05"     "T06"     "T06"     "T07"     "T07"     "T08"
## [176] "T08"     "T09"     "T09"     "T10"     "T10"
```

```
# get Apple x Tree random effects
AT_re <- getME(m, "b")[1:160]
AT_b0 <- AT_re[c(FALSE, TRUE)]
AT_b1 <- AT_re[c(TRUE, FALSE)]
hist(AT_b0, bins = 15, main = "Histogram of Apple x Tree random effects on the intercept")
```

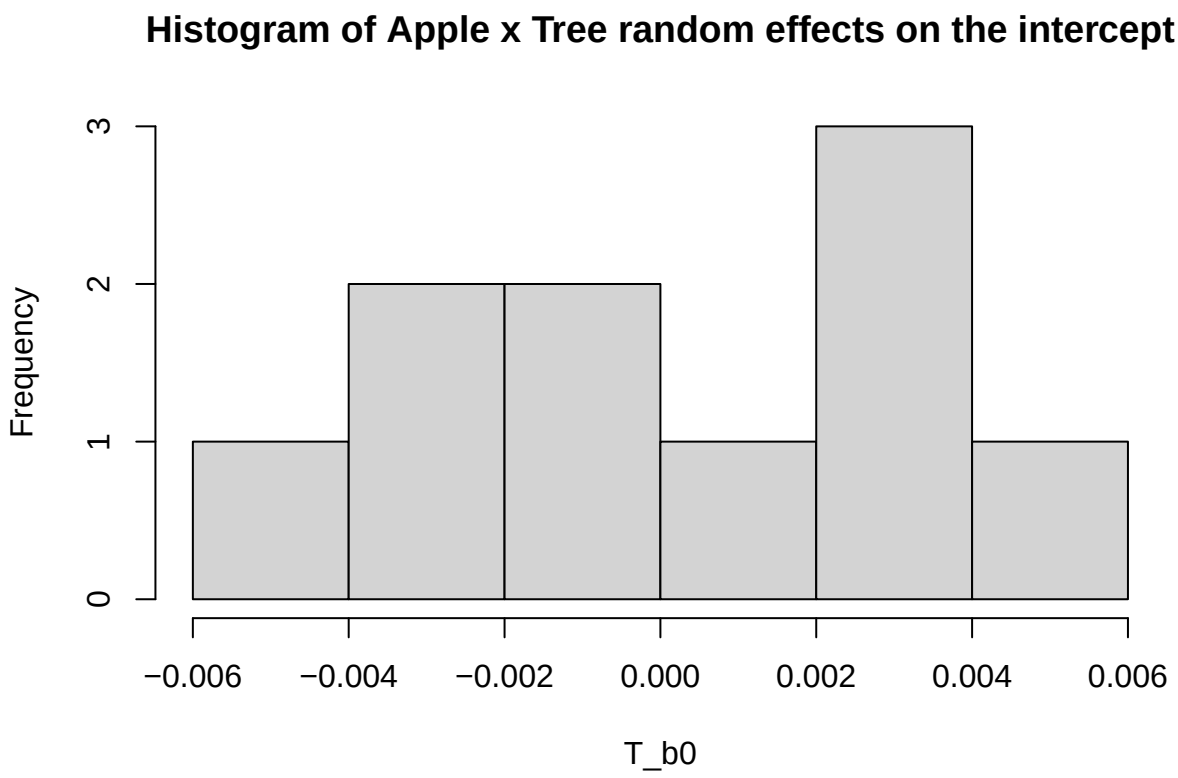



```
hist(AT_b1, bins = 15, main = "Histogram of Apple x Tree random effects on the slope")
```

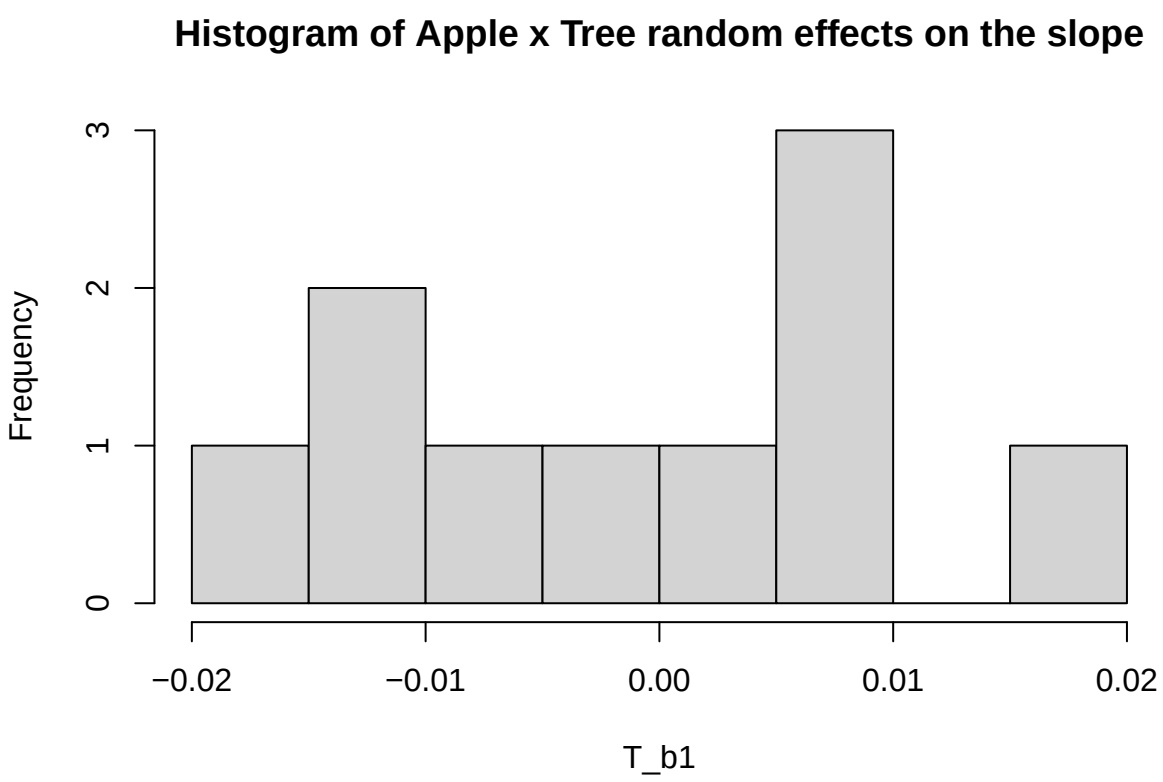



```
# get Tree random effects
T_re <- getME(m, "b")[-c(1:160)]
T_b0 <- T_re[c(FALSE, TRUE)]
T_b1 <- T_re[c(TRUE, FALSE)]

hist(T_b0, bins = 10, main = "Histogram of Apple x Tree random effects on the intercept")
```




```
hist(T_b1, bins = 10, main = "Histogram of Apple x Tree random effects on the slope")
```


Histogram of Apple x Tree random effects on the slope
Frequency
0
1
2
3
−0.02
−0.01
0.00
0.01
0.02
T_b1