Study on fertility of *A. thaliana*

Karim Saied (karim.saied@unil.ch)

November 26th, 2017

1. Question

The aim of the current analysis is to assess the effect of some factors on *Arabidopsis thaliana*'s fertility.

2. Dataset description

A dataset relative to a study on arabidposis' fertility is analysed. The dataset contains 625 observations. The factors assessed are the following:

- `nutrient`: fertilization type. Two-levels factor (`1`: minimal nutrient, `8`: added nutrient)
- `amd`: simulated herbivory (apical meristem damage). Two-levels factor (`clipped`, `unclipped`)
- `gen`: genotype. Twenty-four-levels factor (twenty-four genotypes)
- `reg`: regions indicating where the seeds where collected. Three-levels factor (`NL`: Netherlands, `SP`: Spain, `SW`: Sweden)
- `popu`: population within each region. Nine-levels factor
- `rack`: the rack where the plants were stored on. Two-levels factor (one for each greenhouse racks)
- `status`: the germination method. Three-levels factor (`Normal`, `Petri.Plate`, `Transplant`)
- `total.fruits`: total fruits produced per plant. Response variable (discrete)

Total fruits produced per plant (`total.fruits`) is the measure of fertility. This is the response variable. Knowing that factors `nutrient` and `amd` were declared as being factors of interest, they both are considered as fixed effects whereas `gen`, `reg`, `popu`, `rack` and `status` are random effets. It has been observed that the variables `gen`, `rack` and `nutrient` were numeric. To correct that, these variables have been set as factors. One must note that populations are nested within their region (Table 1) and genotypes are nested in their population (Table 2).

An overview of the distribution of the response variable shows a Poisson distribution (Figure 1).

One has to take into account that some factors are clearly unbalanced.

Regions:

```
##
##  NL  SW  SP
## 116 217 292
```

For instance, within the variable `reg`, only 116 plants come from Netherlands whereas 217 and 292 plants come from Sweden and Spain respectively.

Genotypes:

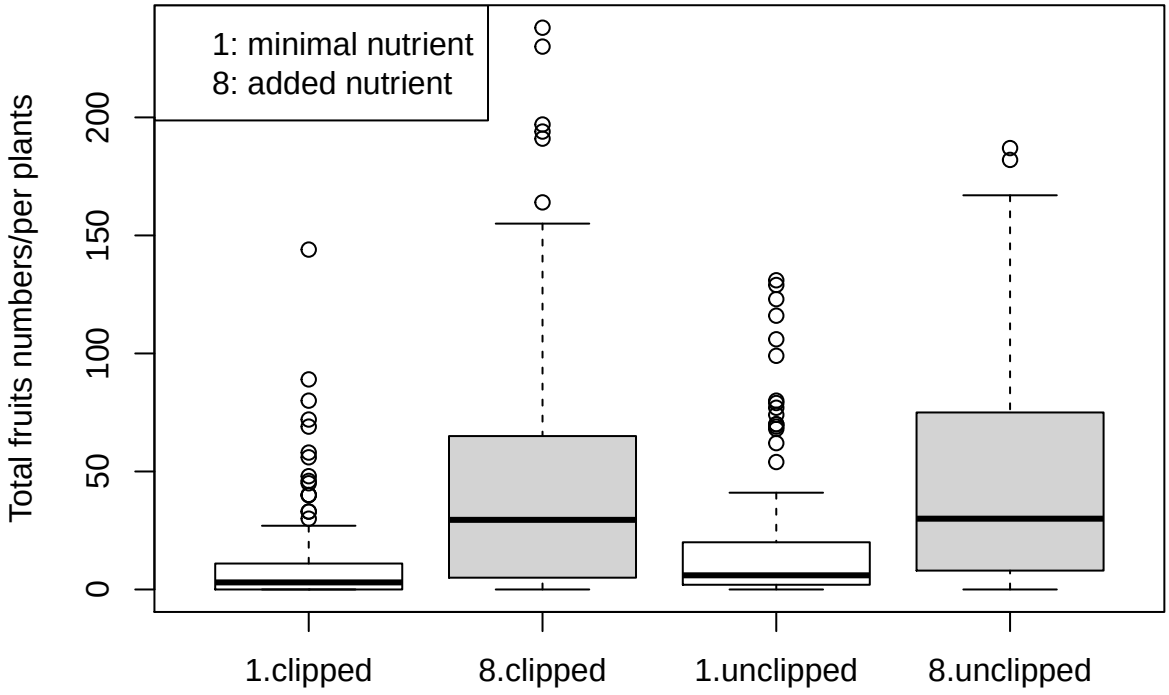
```
##
##  5 18  6 19 22 21 30 23 28 27 17 20 12 14 25  4 11 15 24 13 16 34 36 35
## 11 12 13 13 13 14 14 16 18 20 22 24 26 26 28 31 35 35 35 39 43 45 45 47
```

Another example concerns the variable `gen`. Indeed, the genotype 5 is represented by only 11 plants whereas the genotype 35 is represented by 47 plants.

3. Statistical analysis

As previously mentioned, the response variable `total.fruits` is discrete and fits a Poisson distribution. Moreover, there is a mix of fixed and random effects. Hence, using a generalized linear mixed model fitting a Poisson distribution seems to be a suitable choice.

As a first step, the relashionship between the response variable `total.fruits` and the two fixed effects (`nutrient` and `amd`) is assessed by means of a boxplot. Nevertheless, one has to keep in mind that such visualization only serves to observe trends and nothing can be concluded from it.



Based on that boxplot, plants that are fertilized by adding nutrients (`8`) tend to produce far more fruits than plants with minimal-nutrient level (`1`). That difference is observed within both clipped and unclipped plants. However, simulated herbivory does not seem to affect the amount of fruits produced as many as nutrients do. As a matter of fact, it is shown that for fertilized plants (`8`), the amount of fruits produced is almost similar for clipped and unclipped plants. A very slight difference is still observed. Same observation for plants with a minimal-nutrient level (`1`).

3.1. Generalized linear model

As a second step, the relashionship between fertility and the two fixed effects is assessed using a generalized linear model.

```
##                            Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)              2.3848232 0.02503129  95.27369 0.000000e+00
## nutrient8                1.3999490 0.02778791  50.37978 0.000000e+00
## amdunclipped             0.4336678 0.03158081  13.73200 6.530342e-43
## nutrient8:amdunclipped -0.3684245 0.03570883 -10.31746 5.875521e-25
```

3.2. Generalized linear mixed model

Finally, a generalized linear mixed model is performed in order to take into account both fixed and random effects.

Fixed effects

```
##                            Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)              2.1077763 0.39893144   5.283555 1.267007e-07
## nutrient8                1.4407692 0.02786983  51.696371 0.000000e+00
## amdunclipped             0.4553595 0.03164392  14.390110 5.970099e-47
## nutrient8:amdunclipped -0.4313717 0.03578767 -12.053639 1.855720e-33
```

Regarding the detailed results for the generalized linear model (`glm`) and the generalized linear mixed model (`glmer`), please refer to Table 3 and Table 4 in the *Appendices* section.

4. Discussion and conclusion

First of all, knowing the coefficients tables of `glm` and `glmer` displayed above, one can note that taking into account random effects does not affect estimates so much. However, one observes an improvement of p-values.

Based on the analysis, adding nutrient affect fertility significantly (p-value $<$ 2e-16). Indeed, fertilized plants tend to produce more fruits than plants with minimal-nutrient level. Simulated herbivory (i.e. clipping) also affects significantly fertility (p-value $<$ 2e-16). As a matter of fact, unclipped plants tend to produce slightly more fruits than clipped ones, which indicates that clipping plants negatively affects fruits production. These results are biologically consistent. Indeed, adding nutrients would allow plants to allocate more energy in reproductive function. Furthermore, wounded plants should invest more energy in defense functions at the expense of reproduction which leads to a reduced fertility.

One has to note that the effect of nutrients is 3.2 times higher than the effect of simulated herbivory. This means that nutrients affect far more fruits production than the simulated herbivory. Finally, one can see that there is a significant negative interaction between the two fixed effects (p-value $<$ 2e-16). That interaction is unexpected because it indicates that fertilized unclipped plants tend to have a reduced fertility which is not coherent. That last result should be considered with caution.

In conclusion, unclipped condition and fertilization (i.e. adding nutrients) lead to a higher amount of fruits produced and thus to a higher fertility. However, fertilization seems to be major condition for a higher fertility.

Appendices

Figure 1

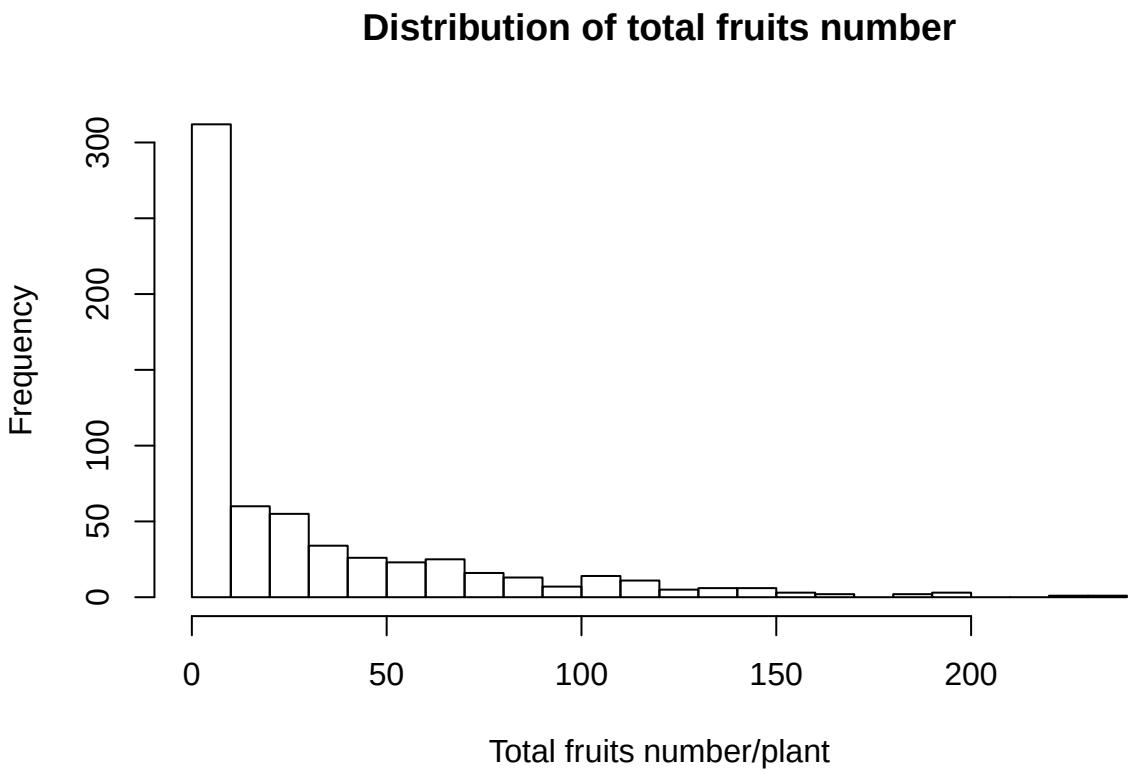


Table 1: populations are nested in their region.

```
##     popu
## reg  1.SP 1.SW 2.SW 3.NL 5.NL 5.SP 6.SP 7.SW 8.SP
##   NL    0    0    0   55   61    0    0    0    0
##   SP  100    0    0    0    0   77   51    0   64
##   SW    0   48   32    0    0    0    0  137    0
```

Table 2: genotypes are nested in populations.

```
##       gen
## popu    4  5  6 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 27 28 30 34
##   1.SP  0  0  0  0  0 39 26 35  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   1.SW  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 28 20  0  0  0
##   2.SW  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 18 14  0
##   3.NL 31 11 13  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   5.NL  0  0  0 35 26  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   5.SP  0  0  0  0  0  0  0  0 43 22 12  0  0  0  0  0  0  0  0  0  0  0
##   6.SP  0  0  0  0  0  0  0  0  0  0  0 13 24 14  0  0  0  0  0  0  0  0
##   7.SW  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 45
##   8.SP  0  0  0  0  0  0  0  0  0  0  0  0  0  0 13 16 35  0  0  0  0  0
##       gen
## popu   35 36
##   1.SP  0  0
##   1.SW  0  0
##   2.SW  0  0
##   3.NL  0  0
##   5.NL  0  0
##   5.SP  0  0
##   6.SP  0  0
##   7.SW 47 45
##   8.SP  0  0
```

As an example, genotypes 13, 14, 15 are only found in population 1 in Spain. Genotypes 20 and 28 are only found in population 1 in Sweden, and so on.

Table 3: detailed results for the generalized linear model (`glm`)

```
##
## Call:
## glm(formula = total.fruits ~ nutrient * amd, family = poisson)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -9.695  -5.086  -2.828   1.818  21.868
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.38482    0.02503   95.27   <2e-16 ***
## nutrient8                1.39995    0.02779   50.38   <2e-16 ***
## amdunclipped             0.43367    0.03158   13.73   <2e-16 ***
## nutrient8:amdunclipped -0.36842    0.03571  -10.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 29269  on 624  degrees of freedom
## Residual deviance: 23550  on 621  degrees of freedom
## AIC: 25921
##
## Number of Fisher Scoring iterations: 6
```

Table 4: detailed results for the generalized linear mixed model (`glmer`)

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: total.fruits ~ nutrient * amd + (1 | gen) + (1 | popu) + (1 |
##     rack) + (1 | reg) + (1 | status)
##
##      AIC      BIC   logLik deviance df.resid
##  18290.8  18330.7  -9136.4  18272.8      616
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -10.123  -3.308  -1.665   1.862  34.075
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  gen    (Intercept) 0.06171  0.2484
##  popu   (Intercept) 0.15469  0.3933
##  status (Intercept) 0.01569  0.1253
##  reg    (Intercept) 0.12149  0.3486
##  rack   (Intercept) 0.18257  0.4273
## Number of obs: 625, groups:  gen, 24; popu, 9; status, 3; reg, 3; rack, 2
##
## Fixed effects:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               2.10778    0.39893    5.28 1.27e-07 ***
## nutrient8                 1.44077    0.02787   51.70  < 2e-16 ***
## amdunclipped              0.45536    0.03164   14.39  < 2e-16 ***
## nutrient8:amdunclipped -0.43137    0.03579  -12.05  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) ntrnt8 amdncl
## nutrient8   -0.056
## amdunclippd -0.050  0.712
## ntrnt8:mdnc  0.044 -0.777 -0.884
```